1.0 STATISTICAL BACKGROUND

1.1 Introduction

In most of the following chapters, it will be assumed that readers have at least taken an introductory course in statistical methods. Some basic concepts will nonetheless be reviewed in this chapter to provide background for the following material. Some essential definitions are listed here. Students should look these up in any introductory statistics textbook, but preferably in a text that they have used in the past. An effort has been made to keep the introductory terminology to a minimum, and it will be supplemented as we go along, and by auxillary reading.

1.2 Some basic statistical concepts

Random variables

In even quite simple situations, we need to be able to distinguish between an abstract label for an observation, and the observations that we actually make in some real-world situation. Statisticians do this by using capital letters (X_1, X_2, X_3, ..., X_N) for the abstract label and lower case letters (x_1, x_2, ..., x_n) for the observations we make in practice. Note that the ellipsis (...) means that some letters are left out -- from the first three given, we can infer that these are X_4, X_5, etc., thru X_{N-1}). More importantly, note that this is a series of finite length -- N random variables in all. In some cases, we need to consider an indefinitely long series of numbers, and write X_1, X_2, X_3, ... to indicate that fact. Also, note that the random variables run from X_1 to X_N, but that the observations end in x_n. This is because we often want to sample a large population and thus only record n of the N possible observations.

Example 1.1 Coin-tossing Consider a simple coin-tossing example. Put 10 coins in a jar, shake well, spill them out and count the number of heads. You will get observations like the following table (note that the individual observation, x_I, is the total number of heads out of 10 coins and that the table is based on 100 tosses of 10 coins):

```
5,5,2,4,3,4,5,6,5,6
4,6,3,7,4,6,5,3,5,4
6,5,5,2,5,5,3,3,6,7
5,8,4,3,4,5,6,5,5,3
5,6,7,5,8,8,7,3,7,7
5,4,6,5,3,6,4,6,5,4
3,3,6,4,7,5,6,6,3,4
6,4,5,6,6,4,3,4,8,3
6,2,8,5,7,4,6,4,5,6
1,6,6,7,5,3,5,6,7,3
```

One can continue this process indefinitely, so we may have to consider an infinite sample space. In many cases, we will be considering finite sample spaces, although we often will not know N. In this case, we do know that N = 100, but if we are considering some natural population over a large area, we likely will not know N, and we may in fact have estimating N as our objective. There is some ambiguity in notation here

in that N can be considered to be a fixed population of the outcomes of 100 tosses, or a sample (n) of the infinite number of possible tosses.

Much of statistical methodology consists of describing the outcomes of "experiments" like coin-tossing, and making inferences about the process that led to the set of observations. Most of the theory underlying statistical methods depends on having a model for the underlying process. Such models are described as probability density functions (abbreviated as pdf). Such a model for the coin-tossing example is the binomial distribution, often written as Bi(n,p) which says that the probability that a randomly obtained observation denoted as x_i takes the value k is:

$$\text{Prob}\{x_i = k\} = f_k = \binom{n}{k} p^k (1-p)^{n-k} \qquad (1.1)$$

where $\binom{n}{k}$ is evaluated as $\frac{n!}{(n-k)!k!}$, in which, for example, 5! (read as "five factorial") is calculated as 5x4x3x2x1 = 120.

This equation gives the pdf for a binomial having n trials (10 in the coin-tossing example). In the example, the random variable can take 11 possible values 0,1,2,3, ..., 10, but in the 100 trials listed above, we observed no zeros and no 9's or 10's. In many practical examples, we won't know the value of p, and want to estimate it from the observed data. If we can somehow establish that it is appropriate to assume the model of eq.(1.1), then we can calculate its expected value, defined as:

$$E(x) = \int_{x=0}^{x=\infty} x f_x dx = \sum_{x=0}^{x=10} x \binom{n}{x} p^x (1-p)^{n-x} \qquad (1.2)$$

since we are here considering a discrete random variable that is only defined on the sample space 0,1,2,...,10, the integral can be replaced with a summation, and this can be evaluated with some algebra to find that E(x) = np. We can then turn this around to estimate p from the mean value of our sample, which is calculated as the sum of the observations (496) divided by the number of observations (100) or E(x) = np = 4.96. Since n=10, we estimate p as:

$$\hat{p} = \frac{4.96}{10} = 0.496.$$

The "hat" over p denotes that it is an estimate of the parameter, p, of the binomial pdf. From the structure of the experiment we can infer that the value of p should be about 0.5, that is, if the coin is "unbiased", the probability that it turns up heads should be 1/2.

The sample mean, $\bar{x} = \Sigma x_i/n$ is often described as a "statistic" derived from a set of observations. Other commonly used statistics are the sample variance:

$$s^2 = \frac{\Sigma(x_i - x)^2}{n-1}$$

and the standard error of the mean, s.e. = $[\frac{s^2}{n}]^{1/2}$. Note that statistics are functions of the data. The mean can be written as $\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n)$, which is a linear function of the random variables $x_1, x_2, ..., x_n$. There are some simple rules from probability theory about linear functions of random variables that make it easy to derive useful results about means.

No doubt the most important probability density function (pdf) in statistics is the normal distribution, which is written as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{\left(2\sigma^2\right)}} \qquad (1.3)$$

The parameter μ is the mean of the distribution and σ the standard deviation. Tables of the frequency distribution (f(x)) of this distribution are available in almost any statistics text, but with parameters $\mu = 0$ and $\sigma = 1$, which is described as the unit normal distribution or standard normal distribution, often represented by the notation N(0,1), while observations drawn from eq.(1.3) are described as $N(\mu,\sigma^2)$.

1.3 The Central Limit Theorem

A very useful result from mathematical statistics is the Central limit theorem:

> **"Let X be a random variable with mean μ and variance σ^2, then the random variable Z:**
>
> $$\mathbf{Z} = \frac{\left(\bar{\mathbf{x}} - \mu\right)\sqrt{\mathbf{n}}}{\sigma}$$
>
> **has a distribution that approaches the standard normal distribution as n approaches infinity.**

This says that, if n is large, then we are virtually guaranteed that the sample mean will have nearly a normal distribution. Inasmuch as the great bulk of modern statistical methods depend on the normal distribution, this result is very reassuring. The important question then is "how large must n be for approximate normality?", and the answer depends very much on the frequency distribution underlying the observed x_i.

Example 1.2 Frequency distributions Consider the data from the coin-tossing experiment (Example 1.1). The random variable tabulated is the number of heads in 10 tosses. We can tabulate the frequency of each outcome (0,1,2,3,...,10 heads) and compare it with the expected frequency calculated from eq.(1.1), giving the following result:

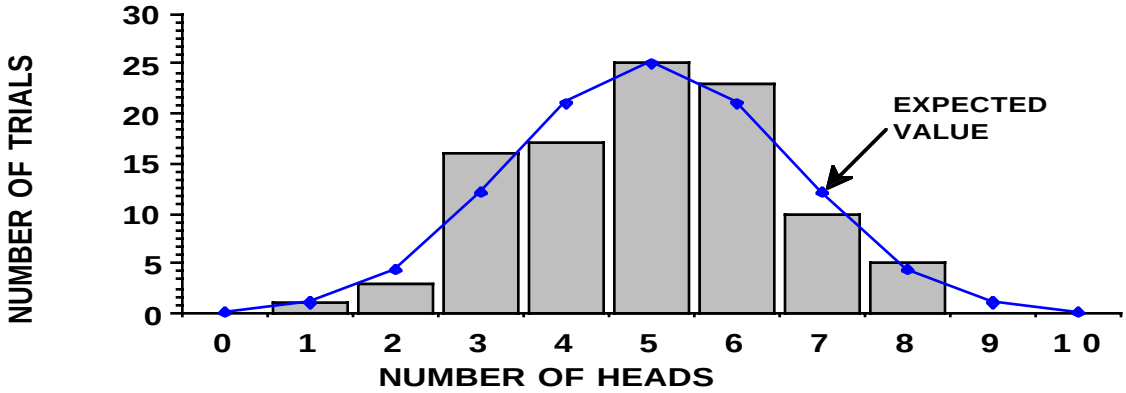


Fig. 1.1 Frequency distribution of number of heads observed in 100 tosses of a coin compared to number expected from eq.(1.1), Bi(10,0.5).

The observed data are not as symmetrical as the expected binomial distribution, but the variance (2.34) is a reasonably good approximation to the variance from the theoretical binomial (2.5) and the mean (4.96) of 100 trials is very close to the theoretically expected value (5). The expected binomial variance of the random variable x, the number of heads in 10 tosses, is readily calculated as np(1-p) = 5(.5)(.5) = 2.5. It is worthwhile to compare (Fig. 1.2) the expected binomial distribution with a normal distribution with the theoretical mean and variance, as calculated from eq.(1.2).

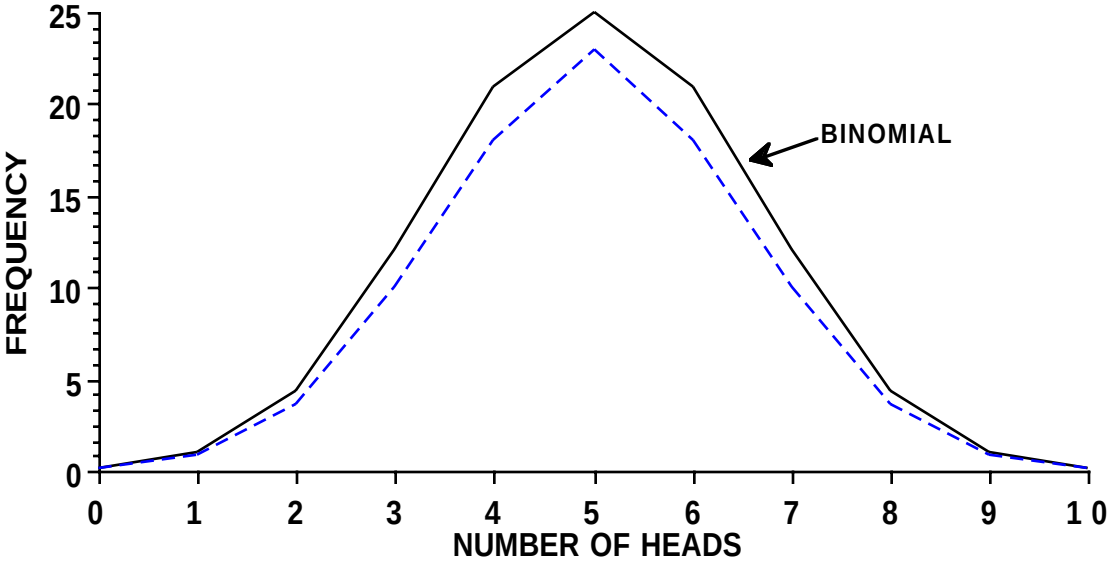


Fig. 1.2 Expected values from a binomial distribution of outcomes of 10 tosses of coins compared to frequencies calculated from a normal distribution (broken line) with the theoretical mean (5) and variance (2.5) for the binomial distribution.

Note that the normal distribution is continuous, i.e. that it takes on all values over the interval considered and is thus only an approximation to the discrete distribution of the results of coin-tossing, in which only integer values can be observed (i = 1,2,3,...,n heads). Hence the points representing the binomial distribution in Fig. 1.1 properly should not be connected by lines. Because the normal distribution has an infinite range it isn't strictly proper to use it in

Fig. 1.2 because there is only a finite possible range of outcomes (0 to 10). However, it is often used as an approximation. Note, too, that there is less area under the normal distribution in Fig. 1.2 because theoretically some observations will be greater and lesser than the range plotted.

1.4 Simple linear regression

Simple linear regression follows the model:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1.4)$$

where y_i is the dependent variable and x_i the independent variable and the error term (ε) is a deviation from the "true" relationship. Estimates of α and β are frequently written as a and b, giving the estimated or fitted relationship as:

$$y_i = a + bx_i \qquad (1.5)$$

Estimates of regression parameters, α and β do not require any assumptions, and can be calculated from any set of x,y pairs. However, tests of significance and confidence limits require adding some assumptions, which center around the ε_i being normally distributed with mean zero and variance σ^2. The assumptions will be discussed after we consider the "machinery" of regression analysis.

The estimates are obtained by the method of least-squares, an important and useful tool that traces back to Legendre and Gauss (known also for the normal distribution) in the early 1800's. Other ways of fitting a straight line to data are available, but seldom used. The approach is based on minimizing a sum of squared deviations, written as:

$$S = \Sigma[y_i - (\alpha + \beta x_i)]^2 \qquad (1.6)$$

where the summation runs from 1 to n. This is accomplished by the methods of calculus, finding the partial derivatives:

$$\frac{\partial S}{\partial \alpha} = 2\Sigma(y_i - \alpha - \beta x_i) = 0 \qquad (1.7)$$

$$\frac{\partial S}{\partial \beta} = 2\Sigma x_i(y_i - \alpha - \beta x_i) = 0$$

these give the normal equations (α and β are replaced by the symbols for estimates, a and b):

$$\Sigma y_i = na + b\Sigma x_i \qquad (1.8)$$
$$\Sigma y_i x_i = a\Sigma x_i + b\Sigma x_i^2$$

and these can be solved jointly to give the estimates:

$$a = \bar{y} - b\bar{x} \qquad (1.9)$$

$$b = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

Note that the deviations of eq.(1.4) are in the vertical plane, being deviations of y_i from the fitted line. Fig. 1.3 shows two of the deviations from a regression line fitted to some counts of deer. The fitted line appears on the graph along with a measure of the fit, R^2, which will be defined below.

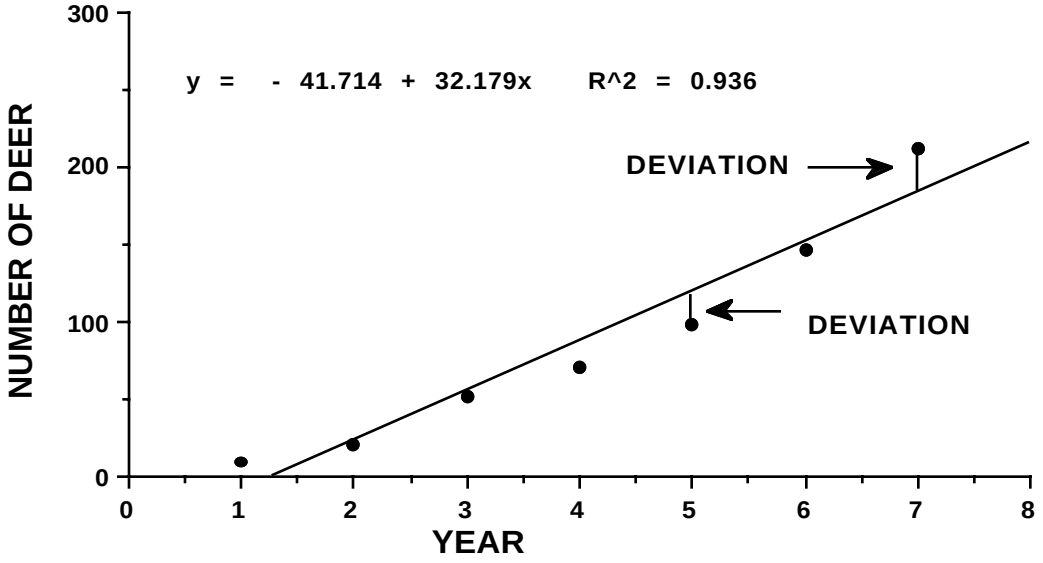


Fig. 1.3. Simple linear regression fitted to successive counts of the number of deer on a study area by the method of least squares.

Table 1.1 gives the analysis of variance results for the deer data from EXCEL, in ANOVA format (the analysis of variance is discussed in Chapter 6). Figure 1.4 shows the deviations from the mean of the y-values, and a comparison with Fig. 1.3 shows why the reduction in Sum of Squares from regression is so substantial (compare Total SS with Residual SS). The residual S.S. is computed from the residuals from the fitted regression line, i.e.:

$$\text{Residual S. S.} = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = \sum_{i=1}^{n} [y_i - ((\bar{y} - b\bar{x}) + bx_i)]^2 \qquad (1.10)$$

$$= \sum_{i=1}^{n} [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

$$= \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1.11)$$

Residual SS = Total SS - Regression SS

Eq.(1.11) can be obtained by introducing the definition of b after squaring the intermediate step above.

Table 1.1 Analysis of variance in regression of deer data of Fig. 1.3 as obtained in EXCEL.

ANOVA

	df	*SS*	*MS*	*F*	*P value*
Regression	1	28992.89	28992.89	73.25	0.0004
Residual	5	1979.11	395.82		
Total	6	30972.00			

	Coefficients	*Standard Error*	*t Statistic*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-41.71	16.81	-2.48	0.06	-84.94	1.51
Slope (b)	32.18	3.76	8.56	0.00	22.51	41.84

EXCEL gives the slope coefficient (b) as "X Variable 1" because the regression program is also set up to handle multiple regression, where there will be 2 or more independent variables. ANOVA is discussed in detail in Chapter 6.

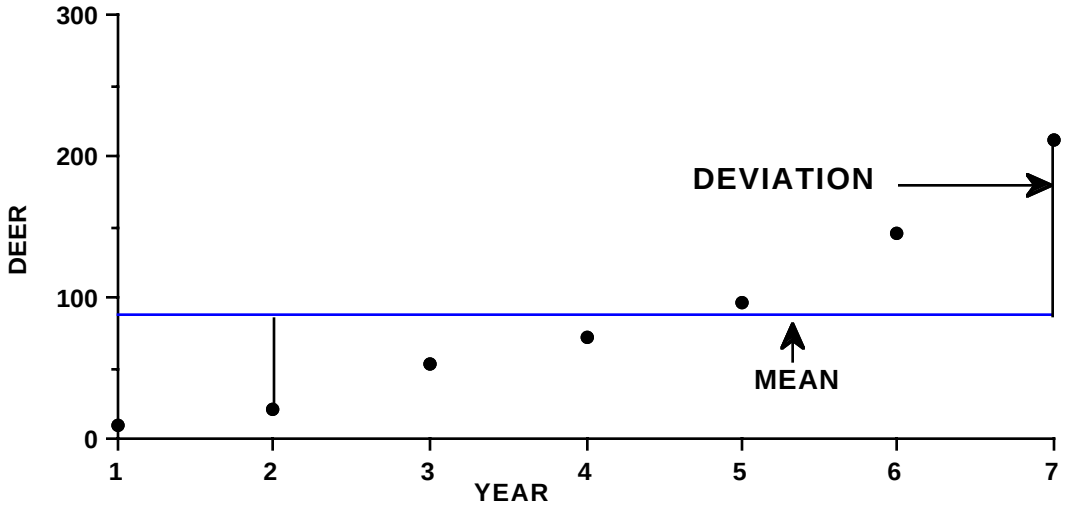


Fig. 1.4 Deer data as in Fig. 1.3 but showing deviations from the mean of the y-values, $\bar{y}$. This shows why the Residual S.S. is ordinarily much smaller than the Total Sum of Squares, which is calculated from the deviations illustrated here.

If the F-value is not significant, there clearly is not much to be gained from the regression line. For simple linear regression, the square root of R-squared (R) is Pearson's product-moment correlation, usually simply referred

to as "the" correlation coefficient (but written as a lower-case r), and calculated as follows:

$$r = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]^{1/2}} \tag{1.12}$$

The correlation coefficient is related to the slope of the regression line (b) by the following expression:

$$b = [\frac{\Sigma(y_i - \bar{y})^2}{\Sigma(x_i - \bar{x})^2}]^{1/2} \; r \tag{1.13}$$

this is sometimes expressed by $\frac{s_y}{s_x}$ r, i.e., the ratio of the sample standard deviation of y to that of x times r. R^2 is also used for multiple regression (described below), where the square root is not the ordinary correlation coefficient, so it is useful to have another expression for R^2. This is:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1.14}$$

The quantity R^2 is often described as measuring the "percent of variance accounted for by regression", in consequence of the fact that it is the ratio of the Regression SS to the Total SS.

Another valuable expression is that of the estimated variance of the slope:

$$s_b^2 = \frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1.15}$$

This expression is particularly useful because it makes it possible to suggest how the estimate of b with smallest variance might be obtained. Concentrating the selection of values of x_i at which to observe y_i at the ends of the possible range of x will evidently give the smallest obtainable variance on b (by giving the largest possible value of the denominator in eq.(1.15)). However, such a course is recommended only when one can be virtually certain that the underlying relationship is linear. We will consider ways to test for nonlinearity in the regression line in a section below. Note, for example, that the data of Fig. 1.3 seem clearly to follow a curved relationship. Concentrating the observations at x-values at the ends of the range of observable y would make it impossible to detect such curvature. Whether we can concentrate observations depends, of course, on the nature of the data. In the case of the counts of deer, we normally make only one observation per year, if the data are an actual census (i.e., a complete count of the deer on an area). In the case of a sample estimate of the number present, it may be

possible to take repeated, independent samples and thus get several estimates per year (replicates).

A confidence interval for the slope, b, uses the t-distribution:

$$b \pm t_{\alpha,d.f.} s_b \qquad (1.16)$$

Note that α now represents the significance level for the t-distribution, and not the parameter of a regression line. Additional confidence intervals for values predicted from the regression line of y or $\bar{y}$ for a given x are given in standard references (e.g., Snedecor and Cochran). Much more detail on regression analysis is given in texts on the subject. An extensive treatment is given by Draper and Smith (Applied Regression Analysis, J. Wiley and Sons Third Edition, 1998). The main parts of the book are presented in matrix algebra notation, but the authors do give a short introduction to the matrix algebra that is adequate to let one follow their presentation of regression topics, and not difficult to understand.

In order to justify any significance tests in regression analysis we must consider the assumptions. The model now becomes:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1.17)$$

where, as with the ANOVA model, we now assume that the ε_i are normally distributed with mean zero, variance σ^2, and are uncorrelated (independent). An important additional assumption is that the x_i values are all measured without error. If the x_i are subject to measurement ("sampling") variation, then the regression line can still be calculated as given above, but its interpretation changes, as do the tests of significance. For the most part, the assumptions for linear regression are somewhat less troublesome than for ANOVA in general. However, we usually need large numbers of replicates to do any testing of the assumptions. Possibly the most important precaution is to be sure that any replicate values of y are indeed obtained independently. In much ecological data it appears likely that the variances of sets of y-values may be proportional to the x_i at which they are taken, or that the coefficients of variation of the replicate y-values may be approximately constant. The F-tests will then be less-reliable. However, simple linear regression is quite "robust" to uncertainties about the assumption of normal errors, so long as the x-values are not subject to error.

A simulation is useful in appraising the assumptions for simple linear regression. Using eq. (1.17) as

$$y_i = 2 + 0.30x_i + \varepsilon_i$$

with the x_i as 1,2,3, ... ,10 and the ε_i generated as observations from a normal distribution with mean 0 and variance $= \sigma^2 = 1$, one can generate a table of "data" as before. This was done to produce a set of data for 20 regression lines. The first 5 data sets are as follows:

		Simulated y_i				
x	True y	1	2	3	4	5
1	2.30	2.86	1.28	2.61	3.26	1.88
2	2.60	0.90	2.08	2.58	1.68	1.11
3	2.90	1.56	3.35	1.35	4.28	3.50
4	3.20	3.85	2.84	2.02	3.67	2.46
5	3.50	1.62	4.20	4.87	1.49	4.68
6	3.80	4.39	5.78	5.22	2.77	3.62
7	4.10	3.66	2.61	4.70	4.24	4.99
8	4.40	3.95	3.90	5.98	2.59	3.81
9	4.70	4.45	6.15	6.41	5.53	3.72
10	5.00	4.50	5.53	6.09	3.54	4.32

Note that the simulated data vary appreciably from the "true values" computed from $y_i = 2 + 0.30x_i$, which appear in the second column above. The simulated data points should follow a normal distribution around the true regression line. Plotting the data (Fig1.5) suggests a certain amount of clumping near the center in some cases, but also shows considerable variability around the true line. If we plot all 200 deviations used to construct the simulations (20 simulations for each of 10 x-values (Fig. 1.6) then it does appear that the underlying distribution is roughly symmetrical, but it should be apparent that one cannot do much testing for normality with smaller samples (say 10 or 20) of deviations from a regression line.

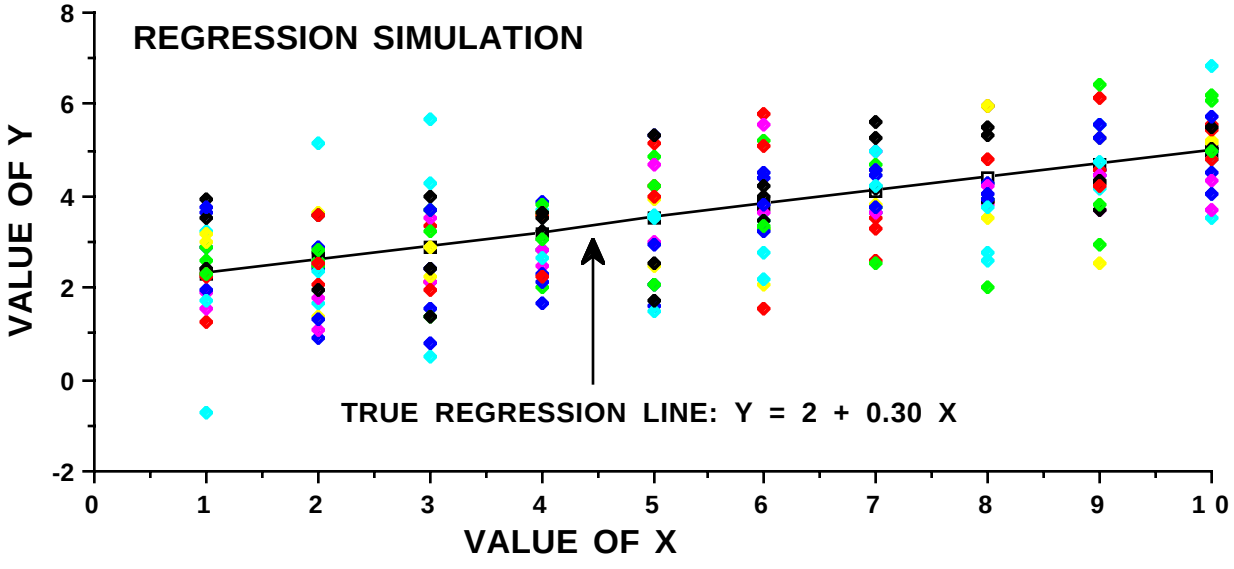


Fig. 1.5 Simulated regression data plotted with the true regression line from which the data were simulated by adding normal deviates with mean zero and unit variance.

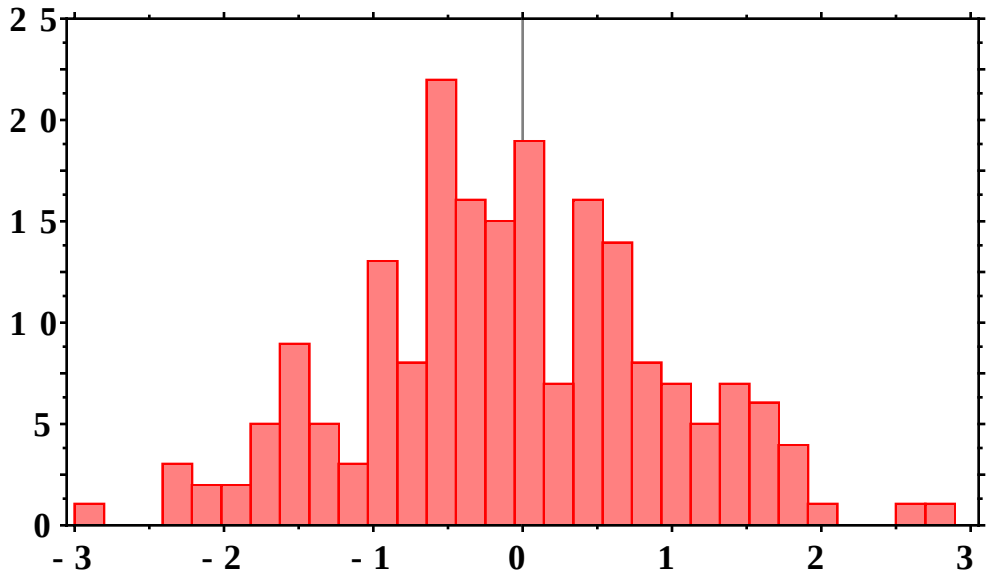


Fig. 1.6 Plot of 200 normal deviations with mean zero and unit variance used to obtain 20 regression simulations.

The regression program in EXCEL used to produce Table 1.1 was run on all 20 sets of generated data and the estimates of intercept (a) and slope (b) were tabulated along with the residual M.S. (s^2) and the confidence limits for b. The error M.S. estimates ranged from 0.5 to 2.41, but averaged 1.05, very close to the expected 1.0. Estimates of the intercept (a; true value 2.0) ranged from 0.2 to 2.83, averaging 2.02, while slope estimates (b; true value 0.30) ranged from 0.11 to 0.54, averaging 0.31. The 95% confidence limits (Fig. 1.7) for the 20 regression estimates of the slope (b) vary considerably, but include the true value in 19 of 20 cases, as expected (0.95(20)=19). It should be noted that this was a fortuitous outcome -- much larger simulations would be needed to be sure that the confidence limits actually include the true β in 95% of cases.

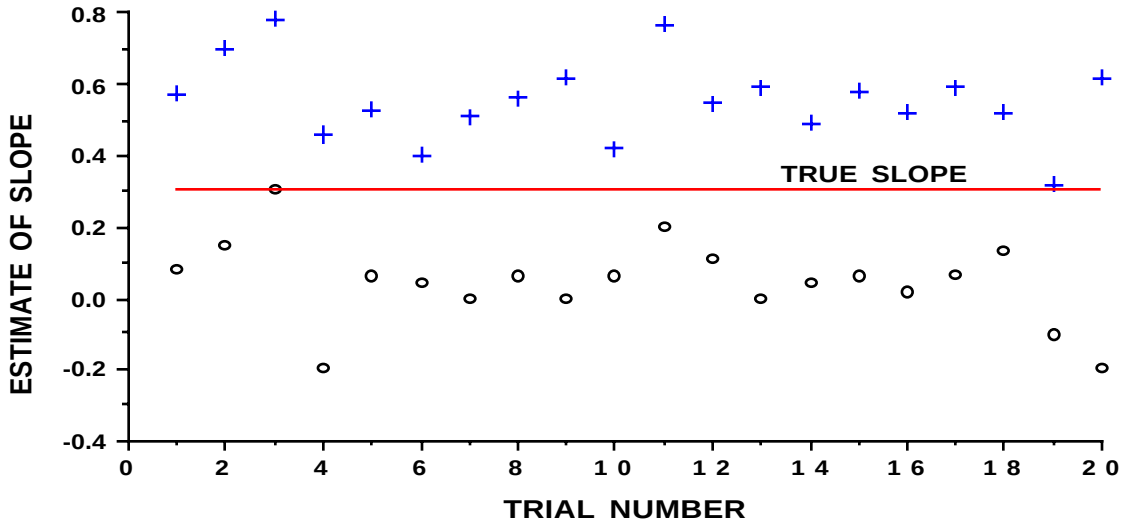


Fig. 1.7 Confidence limits (95%) for slope of 20 simulated regression lines,

shown with the true value (0.30). Note that confidence limits for the 3rd data set do not include (are above) the true value.

1.5 Multiple regression

Multiple regression is somewhat of a risky proposition for ecologists, inasmuch as relationships between several ecological variables tend not to be linear. However, it can be used to explore curvilinear relationships (which we will do below and in the Exercises) and there are various circumstances where a linear model may be useful. It is also true that a multiple regression model is behind many other kinds of analyses. The analysis of variance can be obtained through a multiple regression model, but with a different structure than that used here.

The general model is like that for simple linear regression, but adds more independent variables. We will use 2 here, but EXCEL will compute models with many x-variables. The basic model is:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \tag{1.18}$$

and the same assumptions are made. We again minimize the sum of squares leading to normal equations in three variables and the following solutions for the parameters (a, b_1, b_2):

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \tag{1.19}$$

$$b_1 = [\Sigma(x_{2i} - \bar{x}_2)^2 \Sigma(y_i - \bar{y})(x_{1i} - \bar{x}_1) - \Sigma(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\Sigma(y_i - \bar{y})(x_{2i} - \bar{x}_2)] / D$$

$$b_2 = [(\Sigma(x_{1i} - \bar{x}_1)^2 \Sigma(y_i - \bar{y})(x_{2i} - \bar{x}_2) - \Sigma(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\Sigma(y_i - \bar{y})(x_{1i} - \bar{x}_1)] / D$$

where:

$$D = \Sigma(x_{1i} - \bar{x}_1)^2 \Sigma(x_{2i} - \bar{x}_2)^2 - [\Sigma(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)]^2$$

Our first use of the above equations will be with $x = x_1$ and $x_2 = x^2$, which may look suspicious, but the purpose is legitimate inasmuch as we can now fit a second-degree polynomial (a "quadratic" to many statisticians) as an aid in studying curvature in regression data. To illustrate, we use the deer data of Fig. 1.3 getting the curve of Fig. 1.8. Snedecor and Cochran (1967) show how to do the Analysis of variance in regression in stages, fitting first x_1 and then x_2 to see whether there is any gain in adding a second variable. In the present case, we know that the second variable is necessary to yield a curve.

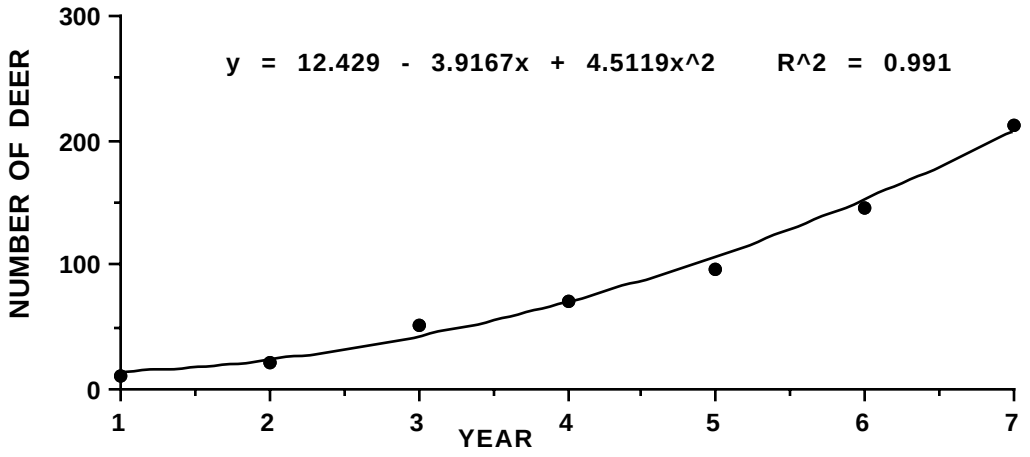


Fig. 1.8. Second degree polynomial fitted to deer data of Fig. 1.3, using multiple regression with $x_1 = x$, and $x_2 = x^2$.

Multiple regression can be used for a wide variety of analyses. For example, the analysis of variance can be represented and computed in a multiple regression format. A wide range of analyses based on multiple regression equations are described in some statistic texts under the heading of "General Linear Hypotheses".

1.6 A test for significant deviations from regression using replicate points.

A test for significant deviations from linearity depending on fitting a curve and testing to see whether the improvement in fit might simply be due to chance will be discussed in the next section. In some cases, however, replicate counts may be available, so that one can use the variability within years to test significance of deviations from linearity. This is the preferred approach, when available. The advantage is that we do not need to specify an alternative model like the quadratic or cubic, which may very well be the wrong model. Note, for example, that population growth data such as that of Fig. 1.8 are known to follow an exponential or geometric curve rather than the second degree polynomial used in Fig. 1.8. Some counts of brown bears at spawning streams provide an example for the test (Fig. 1.9). In this case, the test consists of making the usual analysis of variance to test for significance of the linear regression (Table 1.2), and then using the pooled variance of individual observations within years to estimate "pure error" (Draper and Smith 1998:49). The data for calculation of pooled error appear in Table 1.3. A sum of squares of deviations from the mean is calculated for the data in each year where there are two or more observations and these values are summed to give an overall sum of squares, which is subtracted from the "residual" sum of squares in Table 1.2 to yield the "lack of fit" sum of squares (i.e., the variability not accounted for by "pure error"). The number of counts used to calculate pure error (32) is similarly subtracted from the degrees of freedom for residual error to get the degrees of freedom used to calculate a mean square for "lack of fit". The resulting F-test indicates significance at the 0.05 level, but there does not seem to be much evidence of a consistent pattern of change in Fig. 1.9.

Table 1.2 Test of significance for deviations from regression

	df	*SS*	*MS*	*F*	*Prob.*
Regression	1	0.127	0.127	6.258	0.016
Residual	47	0.954	0.020		
Total	48	1.081			
Lack of fit	15	0.479	0.032	2.150	0.034
Pure error	32	0.475	0.015		

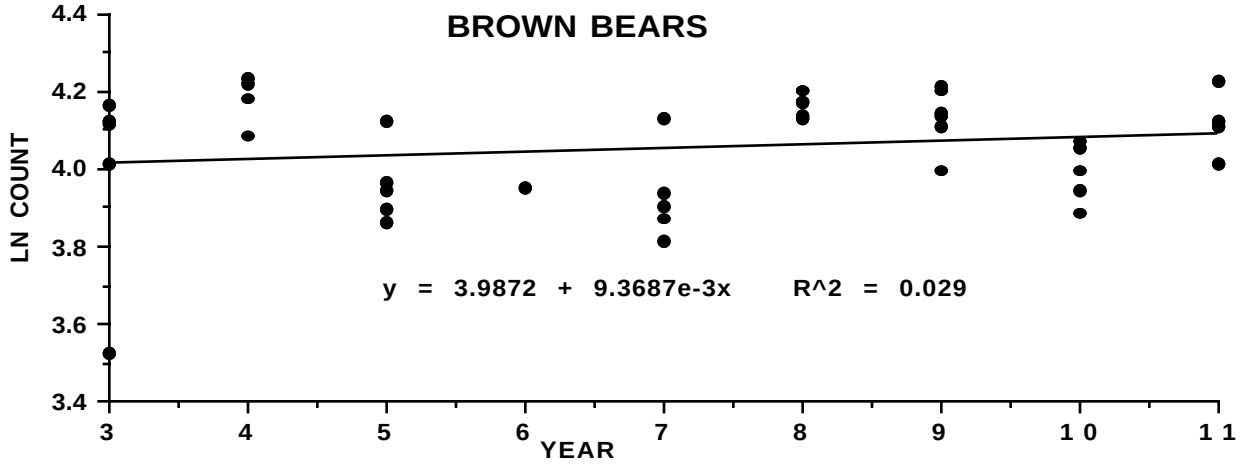


Fig. 1.9 Logarithms of counts of brown bears on salmon spawning streams.

Table 1.3 Data for computation of "pure error" for brown bear counts.

Year	Bears/hour	ln (bears/hr)	Sum of squares	d.f.
3	39.85	9.5219		
3	64.04	4.1595		
3	61.88	4.1252		
3	61.2	4.1141		
3	55.24	4.0117	0.2819	4
4	68.7	4.2297		
4	59.3	4.0826		
4	67.9	4.2180		
4	65.3	4.1790	0.0134	3
5	49.4	9.9000		
5	51.4	9.9396		
5	61.6	4.1207		
5	47.4	9.8586		
5	52.45	9.9599	0.0400	4
6	51.88	9.9489		
7	45.14	9.8098		
7	62	4.1271		

7	48.13	9.8739		
7	49.58	9.9036		
7	51.21	9.9359	0.0572	4
8	62.06	4.1281		
8	66.59	4.1986		
8	62.32	4.1323		
8	66.88	4.2029		
8	65.03	4.1748		
8	64.58	4.1679	0.0051	5
9	54.17	9.9921		
9	67.49	4.2120		
9	66.67	4.1998		
9	62.8	4.1400		
9	61	4.1109		
9	62.42	4.1339	0.0311	5
10	48.68	9.8853		
10	51.47	9.9410		
10	58.51	4.0692		
10	57.65	4.0544		
10	54.08	9.9905	0.0238	4
11	61.12	4.1128		
11	55.15	4.0101		
11	68.29	4.2238		
11	61.52	4.1194	0.0229	3
Sums	2386.08	166.2194	0.4753	32

1.7 Testing for curvilinearity without replications

Trend data are often collected without replications. Occasionally this is because an absolute count is made annually of individuals on an area; more often it is because the investigators cannot afford to make replicate sample counts (seasonal changes limit the time that such "replicates" are likely to be valid, too). In such circumstances, checking for nonlinearity of regression depends on fitting a straight line and a curved line, and appraising the improvement, if any, provided by the curve. The simplest curve available is the second degree polynomial ("quadratic") considered in the section (1.5) on multiple regression above. Sometimes it may be worth trying a third-degree polynomial ("cubic"), which is readily computed by multiple regression in EXCEL. The model is:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \qquad (1.20)$$

where $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$. If a graphics program that fits polynomials is available, it is worthwhile to use it for a quick preliminary check. Often the 3rd degree polynomial has too much curvature, and the practical approach is to stick with the quadratic.

The procedure is straightforward. One first fits the simple linear regression model, obtaining the ANOVA of Table 1.1. Then fit a quadratic, and

obtain the fit illustrated in Fig. 1.8, along with the corresponding regression ANOVA (Table 1.4).

Table 1.4 Analysis of variance in regression based on a multiple regression fit of a second degree polynomial (Fig. 1.8) to the deer data.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	30702.90	15351.45	228.19	0.0001
Residual	4	269.10	67.27		
Total	6	30972.00			

From the linear regression table (Table 1.1), extract the residual sum of squares and use it as the first entry in a new table (Table 1.5). From the ANOVA table giving the multiple regression fit (Table 1.4) also extract the residual sum of squares and make it the second entry in the new table. Use the corresponding degrees of freedom in both cases. Subtract the S. S. for curvilinear regression from the S.S. for linear regression. This quantity, with 1 degree of freedom, represents the improvement in fit provided by curvilinear regression and is tested against the M.S. for curvilinear regression by an F-test. Table 1.5 gives the new arrangement for the deer data.

Table 1.5 Test for curvilinearity of regression using the difference between Residual Sum of Squares in linear regression and multiple regression.

TEST FOR CURVILINEARITY-ORIGINAL SCALE

SOURCE	d.f.	S. S.	M. S.
Dev. from linear regr.	5	1979.11	
Dev. from curvilin. regr	4	269.10	67.27
Difference	1	1710.01	1710.01
F-RATIO		25.42	
SIGNIFICANCE LEVEL		0.0073	

Note that the F-ratio in this table is reversed from the usual regression case. Previously we calculated the F-ratio from $M.S._{regr}/M.S._{resid}$, with 1 and n-2 degrees of freedom. Now we use $M.S._{diff}/M.S._{dev}$ from curvil. regr., with 1 and n-3 d.f. (n-3 because the 3 parameters of eq.(1.19) are fit to the data). In the rare case of using a cubic model, one would use the same procedure, but n-4 d.f. because a 4th parameter is fitted in the cubic (3rd degree polynomial) model.

It is worthwhile to look at the equations for residual S.S. on which the test is based. From eq. (1.10) the S.S. for linear regression is:

$$\text{S. S. Residual (linear regr.)} = \Sigma[y_i - (\alpha + \beta x_i)]^2$$

The corresponding S.S. for the quadratic (2nd degree polynomial) would be:

$$\text{S.S. Residual (quadratic)} = \Sigma[y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i})]^2 \qquad (1.20)$$

From these equations, it can be seen that an improved fit with the quadratic model should reduce the S.S. considerably. If not, then the F-ratio should be small and non-significant.

There are various difficulties in using this test on real data, mostly associated with the inaccuracies of censusing animals and the very real prospect that a population may cease to grow for a variety of reasons. Ecological data are like that! Some statistics books and editors advise checking assumptions before analyses are published. As noted previously here, such tests require more data than are ordinarily available, and may thus be misleading and contradictory.

Following the advice to test assumptions, one might well use the above test to see whether population growth data are linear or non-linear. It is worthwhile to conduct such a test as a way to explore the data. An Exercise asks the student to conduct these tests on actual data on growth of a number of populations. Theoretically, the outcome should be that the test will show nonlinearity and thus lead to using a transformation. In the real world, the results are confusing. The moral is that experience and accepted theory dictate the advisability of a transformation.

1.8 Basic models for population growth

Most ecology textbooks describe population growth by the familiar exponential model:

$$N_t = N_o e^{rt} \qquad (1.22)$$

Where N_t is population size at time t, N_o the starting population size, and r the "instantaneous" rate of population growth. It is worth pointing out that a great many populations do not follow the commonly assumed model, inasmuch as they reproduce only during a short annual period, and thus follow what has been called a "birth-pulse" model, spurting up in numbers at the time of reproduction, and then decreasing through the rest of the year due to mortality. Eq. (1.22) describes continuous change, with reproduction and mortality assumed to be going on constantly in any short time period. A model closer to the truth is of the "compound interest" type:

$$N_t = N_o(1 + r)^t \qquad (1.23)$$

Thus, where equation (1.22) describes a smoothly ascending continous curve, eq.(1.23) describes a "step function" jumping up at specific times and then staying flat in the interim. Neither model is correct at all times, but they do agree at specific times. Figure 1.10 sketches out the likely actual time trend of a population, along with the results of eqs.(1.22) and (1.23). Either model can be described by $N_t = N_o\lambda^t$, with λ representing e^r or (1+r). When we use a log transform to represent population growth data as a straight line (thus performing "log-linear" regression), it is important to have in mind this interpretation of the slope of the regression represented by the two models. Note that eq. (1.23) is actually only defined at the time of reproduction or

recruitment, but the plot (dashed line) connects these “jumps” by a straight line.

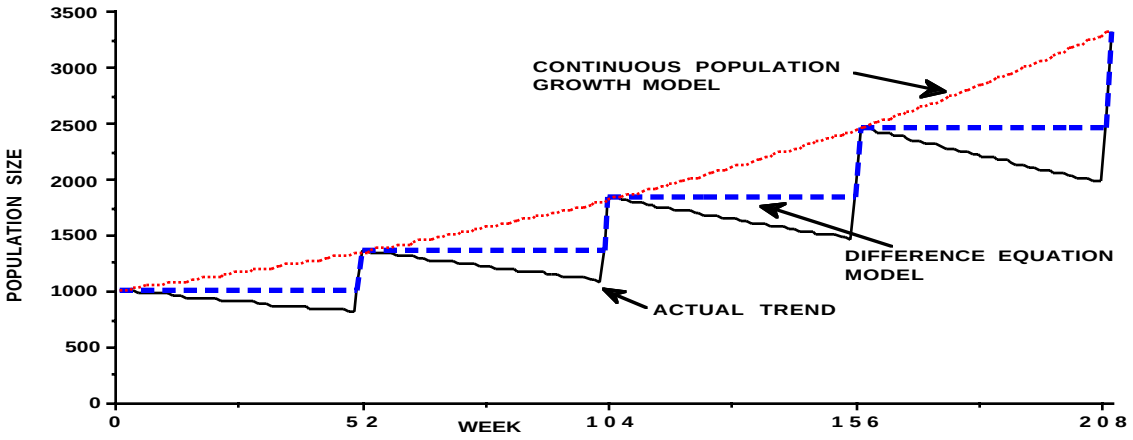


Fig. 1.10 The two population growth models of eqs. (1.22) and (1.23).

1.9 Testing for differences between regression lines

An essential feature of regression analysis is the ability to determine whether a number of fitted regression lines differ. We start out by considering whether the slopes (β_i) of several lines are significantly different. If not, then it is logical to test whether the intercepts (α_i) are different. This leads to the Analysis of Covariance, discussed in the next section.

Most of the data for testing equality of slopes comes from the calculations presented in Section 1.4. The main new feature lies in estimating a common slope. In order to compare the several slopes, we will first need to combine individual slopes to obtain a "pooled" value to compare with the individual values. This also can be obtained by weighting the individual slopes inversely by their variances. The weights come from the variance estimate for individual slopes, eq. (1.15). A basic assumption in assessing regression lines is that they all have the same variance about regression, as estimated by the residual (error) mean square of eq. (1.10). As always, if there is enough data it is worthwhile to test that assumption. Usually only gross differences can be detected with small to moderate sized data sets. If we assume a common variance (s^2), then the weights can be taken as:

$$w_i = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1.24)$$

Thus the slope based on the widest spread of x-values gets the most weight, and the pooled slope becomes:

$$\bar{b} = \frac{\Sigma w_i b_i}{\Sigma w_i} \qquad (1.25)$$

where we have k regression lines to analyze so the summations run from 1 to k. In the analysis, we pool familiar sums of squares for the k regression lines, namely:

$$SSy = \Sigma(y_i - \bar{y})^2, SSx = \Sigma(x_i - \bar{x})^2, and\ SSxy = \Sigma(y_i - \bar{y})(x_i - \bar{x})$$

and use these to arrive at a pooled estimate of the residual (error) sum of squares. The resulting mean square is then used as the denominator in an F-test, where the numerator is:

$$S.S._{diff} = \Sigma w_i(b_i - \bar{b})^2 \qquad (1.26)$$

with k-1 degrees of freedom.

For an example, we compare rates of population increase for data on deer, horses, and elk. Models for rate of growth (eq. (1.22) or (1.23)) indicate that the data should be log-transformed (using logarithms to base e), whereupon the slope of a simple linear regression line will estimate a rate of population growth. This rate of increase for deer (Fig. 1.11) is apparently appreciably higher than those of the other two species. Note that there will be a difference in interpretation of the slopes (b) depending on whether eq.(1.22) or (1.23) is assumed to hold. Details appear in Section 11.2.

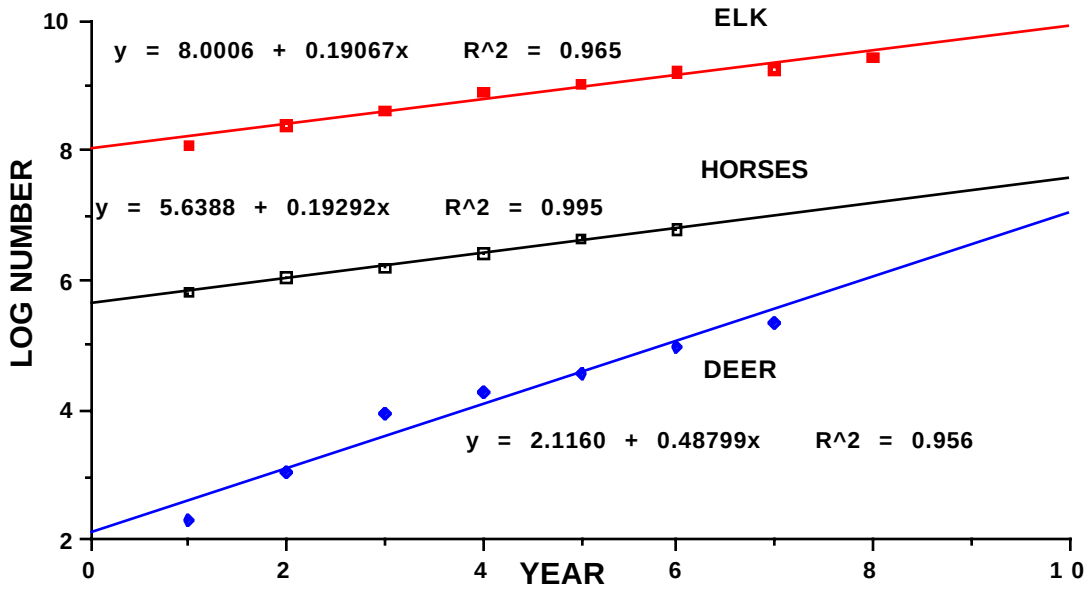


Figure 1.11 Log_e transformed data on numbers of three species with fitted regression lines.

The first 3 columns of data in Table 1.5 are calculated from the individual data sets and summed to get the "pooled" data. The slopes (b_i) are calculated from eq. (1.9), and the first 3 sums of squares (S.S.) on the right are calculated from the right side of eq. (1.11), i.e., from $SSy - b^2SSx$, and summed (totalling 0.363). The fourth S.S. (2.039) in this column is also calculated from eq. (1.11), but using the "pooled" data, while the S.S. labelled "Difference between slopes" (1.676) is obtained as the difference between the pooled value (2.039) and the sum (0.363) of the individual sums of squares. The F-test is the ratio of 2 mean squares, 1.676/0.113 = 14.79 with 1 and 18 d.f., and is

highly significant (P = 0.001) as might be expected from the difference in regression lines (Fig. 1.11).

Table 1.5. Data for a test of significance of equality of slopes for 3 regression lines.

Source	SSx	SSxy	SSy	Slope	d.f.	S.S.	M.S.
Horses	17.5	3.3760	0.6548	0.1929	5	0.00348	0.00070
Deer	28.0	13.6636	6.9713	0.4880	6	0.30359	0.05060
Elk	42.0	8.0081	1.5830	0.1907	7	0.05608	0.00801
						0.36315	
Pooled	87.5	25.048	9.209	0.2863	18	2.03884	0.11327
			Difference between slopes		1	1.676	1.676
						F	14.794
						Prob.	0.0012

The "Difference between slopes" S.S. of Table 1.5 can be calculated directly from eq. (1.26), using the weights calculated from eq. (1.24) to calculate the weighted slope ($\bar{b}$) of eq. (1.25). The calculations appear in Table 1.6. Note that Table 1.6 is not needed for the F-test but provides some further insight into the basis for calculations.

Table 1. 6. Calculations for eq. (1.26).

	Weights	b_i	$w_i b_i$	$(b_i - \bar{b})^2$	$w_i(b_i - \bar{b})^2$
Horse	17.5	0.193	3.376	0.0087	0.1525
Deer	28.0	0.488	13.664	0.0407	1.1394
Elk	42.0	0.191	8.008	0.0091	0.3838
	87.5	Sum	25.048	S.S.	1.6757
		b-bar	0.286		

Another example concerns a situation where it seems likely that the regression intercept (α) should be zero. The data come from a study of Hawaiian monk seals. These seals occupy 5 sites spread over about 1300 miles northwest of the main Hawaiian Islands, and are classified as Endangered under the Endangered Species Act. To monitor their abundance, "beach counts" are conducted annually on most of the sites. These amount to tallying all seals seen in covering all beaches on a site. Only a fraction of the seals using a site are ashore at any given count. However, individual seals can be identified by tags, scar patterns, and the use of temporary bleach marks. In those instances where many counts can be made over 6 weeks or so, it becomes possible to achieve a virtually complete tally of the population using the site. A further description of monk seal dynamics appears in Section 14.5 (Case Histories).

The analysis in this example thus contrasts the mean beach counts against population totals for 3 sites, using regression through the origin. Because α is now assumed zero, the regression model becomes $y_i = \beta x_i + e_i$. The least-squares estimate of β is;

$$b = \frac{\Sigma y_i x_i}{\Sigma x_i^2} \qquad (1.29)$$

which is eq. (1.9) without the means, e.g., $\Sigma(x_i - \bar{x})^2$ is now Σx_i^2. Apart from this change in definitions, the analysis (Table 1. 7) proceeds as in the previous example, with one other exception. Inasmuch as α is not included in the model, we use n-1 d.f. where regression analyses with 2 parameters (α and β) use n-2 d.f.

Table 1. 7 Data for a test of equality of slopes for 3 regression lines relating mean beach counts to total abundance for Hawaiian monk seals at 3 sites.

Source	SSx	SSxy	SSy	Slope	d.f.	S.S.	M.S.
KURE	63431	26670.2	11782.3	0.420	11	568.50	51.68
LAYS	6576511	98198.4	60015.1	0.301	9	283.36	31.48
FFS	1338292	410254.5	126842.6	0.307	4	1078.76	269.69
						1930.62	
Pooled	2059374.0	635123.1	198639.9	0.308	24	2764.196	115.17
	Difference between slopes				1	833.580	833.58
						F	7.238
						Prob.	0.013

It thus appears that there is a significant difference among sites, with one site (Kure) having a significantly greater slope (b) than the other two, where the slopes are virtually identical. The two relationships appear in Fig. 1.12. The site with the largest total counts (French Frigate Shoals) contains many small islands, some of which are small enough that it has been difficult to approach seals for identification. The "total" counts at that site have thus not been considered complete, but the data for the recent 5 years (1991-1995) considered here now suggest that the apparent total counts do agree with the relationship between beach counts and totals at Laysan, suggesting that the FFS data may now approximate actual totals.

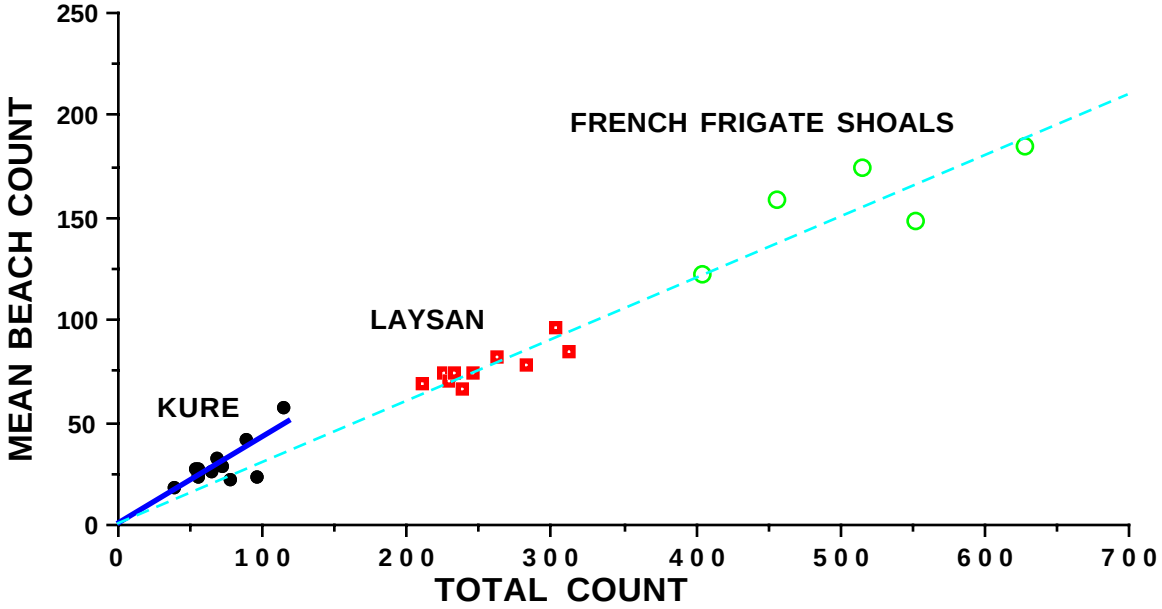


Fig. 1.12. Relationship between mean beach counts and total counts at three monk seal population sites. Regression through the origin for Kure is shown

by a solid line, while the same regression for Laysan and French Frigate Shoals appears as a broken line.

One other issue illustrated by the monk seal data should be discussed here. This is the aggravating question of "outliers". In some data sets, there are points that seem evidently to lie well away from a trend evident in the bulk of the points. This is the case with the Laysan data. There are two years (Fig. 1.13) that are well away from the trend line (and were not used in the analysis of Table 1. 7). Simple and direct methods are not available for deciding to exclude "outliers". However, in extreme cases like this one, we can simply consider the probability of such a deviation. The standard deviation of the distribution of points around the regression line for Laysan is the square root of the Mean Square of Table 1.7, which is $31.48^{1/2} = 5.6$. Deviations of the two suspect points from the regression line are 65 and 62 units, or about 10 standard deviations away from the line. Clearly these two deviations have an extremely low probability of arising by chance alone. Corroboration is also available in that the two points (they occurred in successive years) represent an increase in population size that is simply not feasible, and a subsequent decrease that surely would have been detected (dead seals) if it occurred.

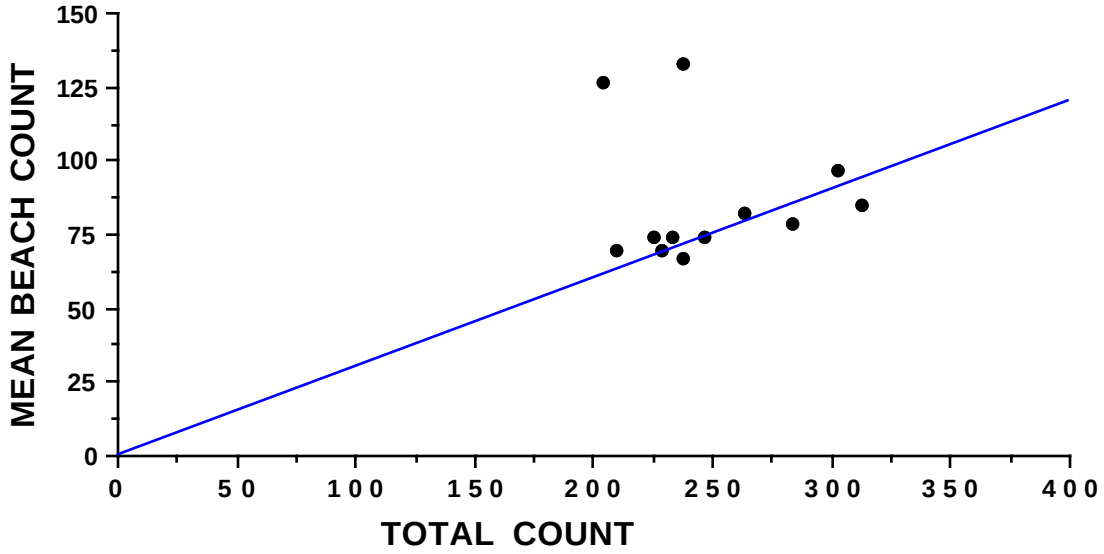


Fig. 1.13. Position of two aberrant counts at Laysan Island relative to the regression line and data from which it was calculated.

1.10 The Analysis of Covariance

The analysis of covariance depends on the availability of an auxiliary measurement linearly related to the variable of interest. Consider a one-way analysis of the yield (y_i) of fruit trees subjected to several different treatments (different types of fertilizer or perhaps insecticides) that presumably will increase yield. Yield of individual trees may vary with the size and location of the tree, so a useful auxiliary variable may be the yield (x_{ij}) of a given tree in the year before the treatments were applied. Hence, a one-way model without information from the auxiliary variable is:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

but the auxiliary variable can be introduced by:

$$y_{ij} = \mu_i + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij} \qquad (1.28)$$

so that the adjusted mean for a given treatment becomes:

$$\bar{y}_{i.} = \mu_i + \beta(\bar{x}_{i.} - \bar{x}_{..}) + \bar{e}_{i.}$$

The ANOVA for a covariance adjustment then tests whether adjusted means are significantly different. The dot notation is used with multiple subscripts to indicate which subscript is involved in averaging. Thus $\bar{x}_{i.}$ Denotes the average over j for the ith group.

A key assumption in the analysis of covariance is that the same linear relationship holds in all of the treatment groups. Thus we need to use the methodology of Section 1.9 to test the hypothesis that β_i within treatment groups are not significantly different. Some investigators may proceed with the analysis without testing homogeneity of the slopes. This is not wise unless there is a good deal of prior experience on which to base such a decision. Inasmuch as both analyses depend on much the same computations, prudence calls for computing the results given in Table 1.5 and 1.7 in any case.

The data are arranged in the same way as in the previous section, but we here assume the same number of observations in each treatment group, giving a table like the following:

A		B		C	
x	y	x	y	x	y
x_{11}	y_{11}	x_{21}	y_{21}	x_{31}	y_{31}
x_{12}	y_{12}	x_{21}	y_{21}	x_{32}	y_{32}
.	.	.	.	.	.
x_{1j}	y_{1j}	x_{2j}	y_{2j}	x_{3j}	y_{3j}
.	.	.	.	.	.
x_{1n}	y_{1n}	x_{2n}	y_{2n}	x_{3n}	y_{3n}

To provide an example, a table of data from Snedecor and Cochran follows:

A		D		F	
x	y	x	y	x	y
11	6	6	0	16	13
8	0	6	2	13	10
5	2	7	3	11	18
14	8	8	1	9	5
19	11	18	18	21	23
6	4	8	4	16	12
10	13	19	14	12	5
6	1	8	9	12	16

	11	8	5	1	7	1
	3	0	15	9	12	20
Means	9.3	5.3	10	6.1	12.9	12.3

From the ANOVA for simple regression we had the following results (eq.(1.11):

$$\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Total S.S. - Regression S.S. = Error (Residual S. S.)

The error term can be written in various ways

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2&=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2\\&=\sum_{i=1}^{n}(y_i-\bar{y})^2-b^2\sum_{i=1}^{n}(x_i-\bar{x})^2\\&=\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{[\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\end{aligned}\qquad(1.29)$$

with the last result being most useful here. It is obtained by using the definition of b in developing eq. (1.29) from equation for the Residual (Error) Sum of Squares above. The above calculations are expressed for one group of data, so in dealing with several groups below, a subscript for the j^{th} observation in the i^{th} group needs to be added.

The calculations proceed by computing the 3 components of eq.(1.29) and arranging them in an ANOVA type of table in which the Total S.S. is calculated from the entire set of data, using overall means of x and y, e.g.with

$$SSx=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{..})^2$$

the other values SSy, and SSxy calculated in the same manner. Thus,

$$SSy=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2,\ \text{and}\ SSxy=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})(x_{ij}-\bar{x}_{..}).$$

The Error line is calculated by using the group means, e.g.,

$$SSx=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i.})^2$$

The Between S.S. are readily obtained by subtracting the line for Error S.S. from Total S. S. These calculations then give the following table from the data above:

Source	d.f.	SSx	SSxy	SSy	M.S.
Total	29	665.9	731.2	1288.7	
Between	2	72.9	145.8	293.6	
Error	27	593.0	585.4	995.1	36.86

The M.S. due to error is calculated from SSy/d.f. = 995.1/27 = 36.86 in the Error line of S.S. just as it would be done without the auxiliary variable. The other entries in the table are needed to obtain a reduction in the error sum of squares as shown below.

The "reduction due to regression" is obtained from $(SSxy)^2/SSx$ in the Error line, and is subtracted from the Error sum of squares as computed without the auxiliary variable, giving an estimate of error mean square adjusted by the regression data. The complete calculation of an adjusted error mean square is thus:

Source	d.f.	SSx	SSxy	SSy	M.S.
Total	29	665.9	731.2	1288.7	
Between	2	72.9	145.8	293.6	
Error	27	593.0	585.4	995.1	36.86
Reduct. due to regr	1			577.9	
Dev. from regr	26			417.2	16.05

An estimate of a common slope is also obtained from the error line, $b = SSxy/SSx^2 = 585.4/593.0 = 0.987$. This value then can be used to get adjusted values of $\bar{y}$ from the following:

$$\bar{y}_{i,adj} = \bar{y}_{i.} - b(\bar{x}_{i.} - \bar{x}_{..})$$

The adjusted mean for the first group of data (group A in the table above) is thus:

$$5.3 - 0.987(9.3 - 10.73) = 6.71 = \bar{y}_{i,adj}$$

The results of the covariance adjustment can then be assembled to produce a covariance-adjusted F-test, as in the following table:

Table 1.8 Covariance F-test in one-way classification

					Deviations from regression			
	d.f.	SSx	SSxy	SSy	Reduc.	d.f.	S.S.	M.S.
Treatments	2	72.867	145.8	293.6				
Error	27	593.000	585.4	995.1	577.9	26	417	16.05
T+E	29	665.867	731.2	1288.7	802.94	28	486	
						2	68.6	34.28

The F-ratio is 34.28/16.05 = 2.14 with 2 and 26 d.f. and does not suggest a significant treatment effect (P = 0.14).

The whole purpose of the exercise is to get a more sensitive F-test of main effects than would be possible without the auxiliary variable. Such an improvement depends, of course, on the presence of a significant linear

relationship between the variable of interest (y_i) and the auxiliary variable (x_i), and this relationship needs to be checked out first (i.e., do regressions on the data in each group (A, D, and F) first).

1.11 ANOVA as a regression model

To sketch out a basis for doing an analysis of variance with a regression model, we need the concept of a "dummy variable" which is simply a variable that takes only values of 0 or 1. Consider the multiple regression model:

$$y_i = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

and let $x_i = 1$ if y_i belongs to a particular group in a one-way ANOVA and 0 otherwise. Then we can write:

$y_1 = \mu + \beta_1$
$y_2 = \mu + \beta_1$
$y_3 = \mu + \beta_1$
$y_4 = \mu + \beta_2$
$y_5 = \mu + \beta_2$
$y_6 = \mu + \beta_2$
$y_7 = \mu + \beta_3$
$y_8 = \mu + \beta_3$
$y_9 = \mu + \beta_3$
$y_{10} = \mu + \beta_4$
$y_{11} = \mu + \beta_4$
$y_{12} = \mu + \beta_4$

and thus have a regression model conforming to a one-way ANOVA with three observations in each of 4 groups, giving the general model of $E(y_i) = \mu + \beta_i$, as is appropriate for one-way analysis of variance. Draper and Smith (1998) give extensions to two-way and higher analysis and methods of fitting. The approach is likely not of much importance here, but is mentioned to emphasize an earlier remark that models of the multiple regression type can be used for a wide variety of purposes, often subsumed under the heading of "General Linear Hypotheses".

1.12 Stepwise regression

This is an approach to regression that permits adding variables one step at a time while searching for the "best" model for a given data set. Consider the test for curvilinearity of Section 1.7. We first fitted a linear regression of the form $y_i = \alpha + \beta_1 x_{1i}$ and then extended the model to become a second degree polynomial $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}^2$, using multiple regression to fit the model. We then tested for a significant "improvement of fit" by comparing the reduction in Sum of Squares obtained by subtracting the deviations from curvilinear regression (Residual S.S.) from the deviations from linear regression, and tested significance of the improvement by an F-test. We noted that the process

could be extended to a third-degree polynomial $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3$ to test for a more extreme curvature. We used multiple regression to fit the models, letting $x_{2i} = x_{1i}^2$, (and $x_{3i} = x_{1i}^3$ if the model were extended to test the further improvement of adding a "cubic" term). This kind of procedure is employed in stepwise regression, but is not, of course, restricted to polynomials. Any series of variables can be tested successively for the improvement of fit produced as each new variable is introduced. Computer programs are available that will test all combinations of a set of candidate variables but the results are practically guaranteed to be misleading, as enough manipulation will almost always produce a "good fit". One should use stepwise regression only when there is a logical sequence of models to test, and even then it is likely that the final model will be "over-fit" (i.e., have too many independent variables). One useful approach is to develop a model on half the data and check it on the other half. Usually, ecologists do not have enough data to hold half of it in reserve while studying a model. An alternative is known as "cross-validation". In it a series of fits are used and each observation is left out in turn, and used to check the error variance estimate from the fitted model. Such a test is "computer-intensive", i.e., depends on the ability of the modern computer to conduct many calculations in a short time. Anyone planning to use stepwise regression should consult references like Draper and Smith (1998) first.

1.13 Logistic regression

This is a form of regression analysis developed for data of the binomial form, i.e., in which the variable of interest is either 1 or 0 (or "yes" or "no", "present or absent", etc., which can be coded as 1 or 0). Usually we express results as a proportion, e.g, the proportion surviving after some time interval or some treatment. Logistic regression originated in the field of bioassay, in which the response to a given dose of some substance is studied quantitatively. If one plots the response (proportion surviving or otherwise responding to some treatment) against the dose (often quantity of some substance given an individual) the resulting curve is usually sigmoid (s-shaped). The cumulative normal curve provides a convenient s-shaped model, and is used in bioassay in "probit" analysis. Details of methods used for bioassay are given by D.J. Finney (Statistical Method in Biological Assay, 3rd Ed. 1978, Charles Griffin and co., Ltd. London).

Joseph Berkson proposed using the logistic function as a bioassay model in 1944. The basic model is:

$$P = \frac{1}{1 + e^{-(a+bx)}} \qquad (1.30)$$

where P denotes the dependent variable and x is the independent variable ("dose" in bioassay). Because P is a proportion,

$$Q = 1 - P = \frac{e^{-(a+bx)}}{1 + e^{-(a+bx)}} \qquad (1.31)$$

and we can now consider the ratio of P and Q:

$$\frac{P}{Q} = e^{\alpha + \beta x} \qquad (1.32)$$

The ratio of P to Q is sometimes called the "odds ratio", no doubt because it expresses the odds for a particular outcome.

Now the natural logarithm of this “odds ratio” {eq.(1.32)]is a linear function,

$$\ln(\frac{P}{Q}) = \alpha + \beta x \qquad (1.33)$$

This is called the "logit" transformation.

There is an interesting sidelight to the logit transform. Consider a table of proportions (e.g., several species of plants classified by whether they have flowers, fruits or neither). One can then calculate the natural logarithm of the "odds ratio" and analyze the linear model of eq. (1.33). This is termed log-linear regression by some authors and can be extended to behave like the analysis of variance. It has been used largely in the social sciences, but could well be of interest in ecological circumstances where one must analyze tables of proportions (or tables in general, for that matter). It should be noted that we will also use the term “loglinear regression” to refer to the log transform of eq.(1.22).

Example 1.3 An example of logistic regression

In aerial counts of wildlife populations, the number of individuals in a group has a marked effect on visibility. This has been studied by using animals with attached radiotransmitters and recording the frequency of observation of groups containing these individuals. Such a study of elk has been used to correct for visibility (M. D. Samuel et al. 1987. Visibility bias during aerial surveys of elk in northcentral Idaho. Journal of Wildlife Management 51:622-630). The following table shows the data (only small samples were available so that larger groups had to be combined).

Table 1.9 Sighting data from an aerial survey of radio-marked elk.

Group size	Missed	Seen	Proportion seen	Logit transformation $\log_e(P/Q)$
1	18	5	0.217	-1.281
2	7	6	0.462	-0.154
3	5	5	0.500	0.000
4	4	6	0.600	0.405
5	4	9	0.692	0.811
6	6	4	0.400	-0.405
11	3	14	0.824	1.540
23	0	10	1.000	

The simplest way to fit this data is to use eq. (1.33), i.e., regress the logit values (right-hand column) against x. In this case, the investigators used the logarithm of group size in their analysis, so we use ln (group size) for x in Fig. 1.14, which shows the regression fit.

Due to the fact that the independent variable is from a binomial distribution the linear model implied by eq. (1.33) does not give the best fit to the data. Instead, the technique of maximum likelihood estimation is recommended. If we assume a particular frequency distribution (probability distribution function in Section 1.2) underlies a set of observations, then it may be possible to find expressions that often minimize the variance of an estimated quantity. Methods of mathematical statistics are required to derive such estimators, but many of the commonly-used estimates are also maximum likelihood estimates. In the present case, there is no simple expression for estimating the parameters of eq. (1.32) so that an iteritive method is required to solve the maximum likelihood equations. The method used here is due to J. Berkson (Tables for the maximum likelihood estimate of the logistic function. Biometrics 13:28-34, 1957). Maximum likelihood estimates for logistic regression can also be obtained in some of the available statistics programs (e.g., SYSTAT).

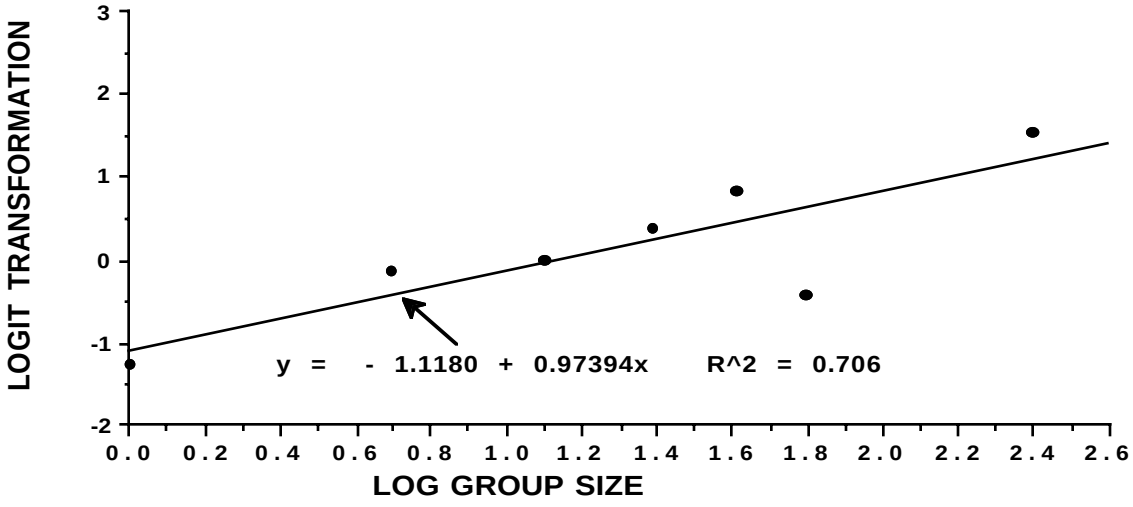


Fig. 1.14. Regression of logit values on logarithm of group size from aerial survey of elk.

The parameters obtained from the regression analysis (Fig. 1.14) are α = -1.118 and β = 0.974, while those obtained from the maximum likelihood fit are somewhat different, being α = -1.305 and β = 1.155. Fits to eq. (1.30) are not substantially different (Fig. 1.15).

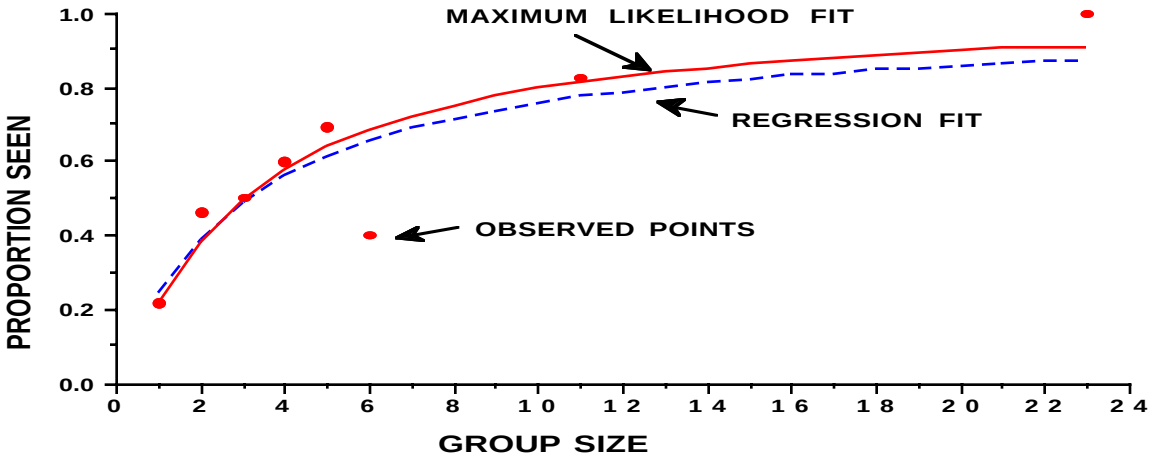


Fig. 1.15 Fits of eq. (1.30) to observed data on elk sightability using regression (eq.(1.33) and maximum likelihood methods.

Example 1.4

Two further examples (Fig. 1.16) are based on reproductive rates in Hawaiian monk seals at two sites. The curves were fitted as above, using regression and maximum likelihood estimates.

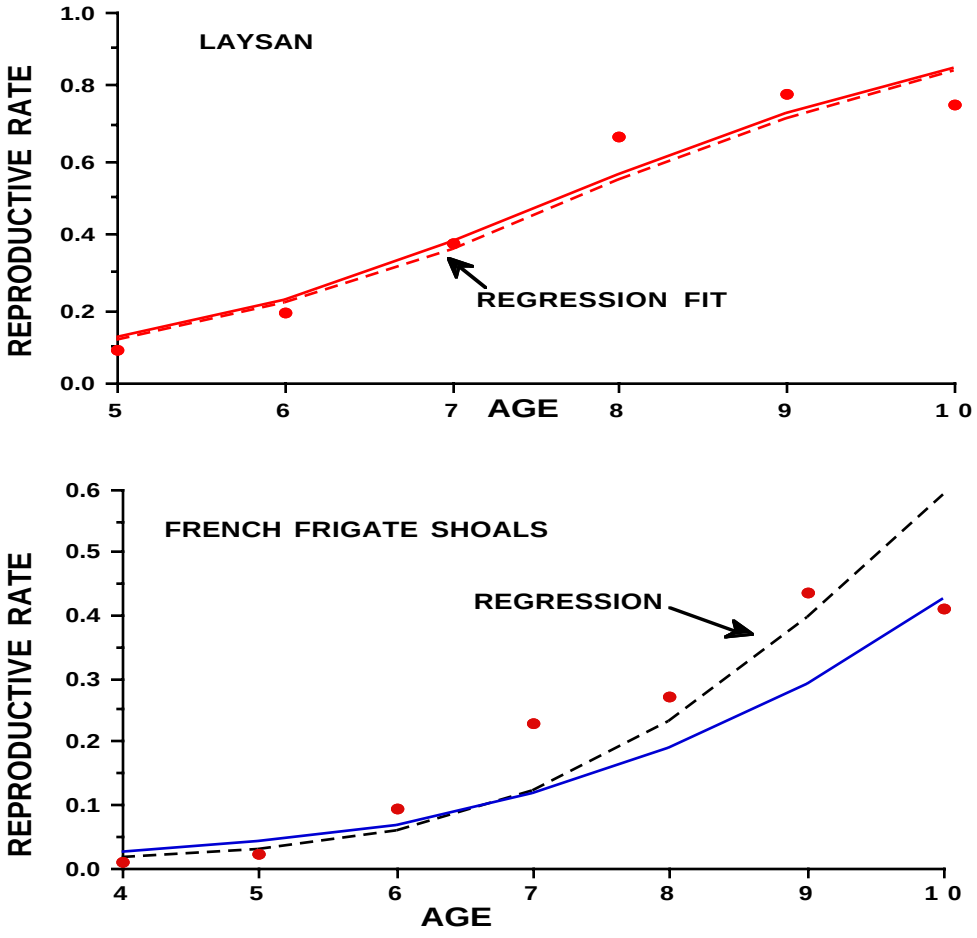


Fig. 1.16 Logistic fits of reproductive rates against age of the female for Hawaiian monk seals at two sites.

In the upper curve, it appears that the regression and maximum likelihood methods give about the same results, while neither provides much of a fit in the lower curve. Deteriorating conditions (poor food supplies and survival) at the site may be changing the curve, so that it does not represent a stable situation. Circumstances at the site shown in the upper curve have been reasonably good, but there is no particular reason to suppose that reproductive rates should follow a logistic curve.

For comparison, some data on judging sound intensity were fitted by the two methods (Fig. 1.17). These data appear to fit the logistic very well, and the two methods of estimation give virtually indistinguishable results.

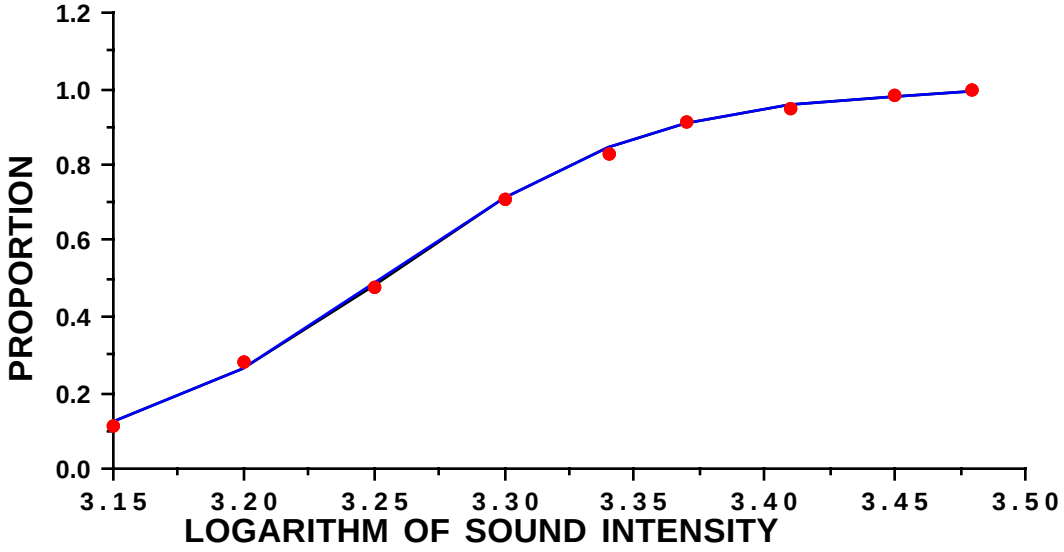


Fig. 1.17. Logistic curve fitted to data on judging sound intensity.

1.14 Locally weighted regression

When there is no suitable model for a curve, locally weighted regression provides a way to fit a smoothed line. The method is variously called "loess" or "lowess". Some authors use "loess", but ecologists will no doubt be confused by the implication of wind-deposited soil!. Weighted linear regressions are fit at each point on the graph (e.g., if the data span 30 years, then such regressions are fit at each of the 30 years) by selecting data points in the immediate neighborhood of each point on the x-abcissa. The number of points in each such neighborhood might be taken to be, say, about 30% of the total number of observations. However, this can be varied in the fitting program, and depends on the purpose at hand. If one wants a thorough smoothing, then 50% or more of the points might be used in each regression. If the smoothed curve is to follow the data point closely, then a small fraction, perhaps as little as 10%, of the points should be used in each fitting. Experimentation with the fitting program will help in developing an approach for a particular data set. Weights diminish by a cubic function, so points very near to the selected point get by far the most weight. The fitted regression line determines only the y-value for the selected abcissal value. In effect, the technique behaves much like a moving average, but has various

advantages. Programs to produce lowess fits are available. SYSTAT has a routine for lowess fitting in the plotting routine (after loading the data in a file, bring up "plot", and select the "smooth" function. It will then be necessary to indicate the fraction of the data points to use in each neighborhood). The lowess method was developed by W. S. Cleveland (Journal Amer. Statistical Assoc. 74:829-836).

The smoothed line in Fig. 1.18 illustrate the technique. This approach to smoothing is preferable to the usual moving-average smoothing because it does not leave blanks at the end of the series, and uses what seems to be a better averaging approach. The lowess technique can be illustrated by smoothing French Frigate Shoals monk seal beach count data. At each point along the line (here, each year) the nearest n points are used to form a weighted linear regression (9 points were used in producing Fig. 1.18). The regression line is used only to determine the smoothed value for the given point. Inasmuch as the weights and the regression line must be computed for every point used along the x-axis (the years 1957 to 1993 in the present example), enough calculations are involved to make use of a computer desirable.

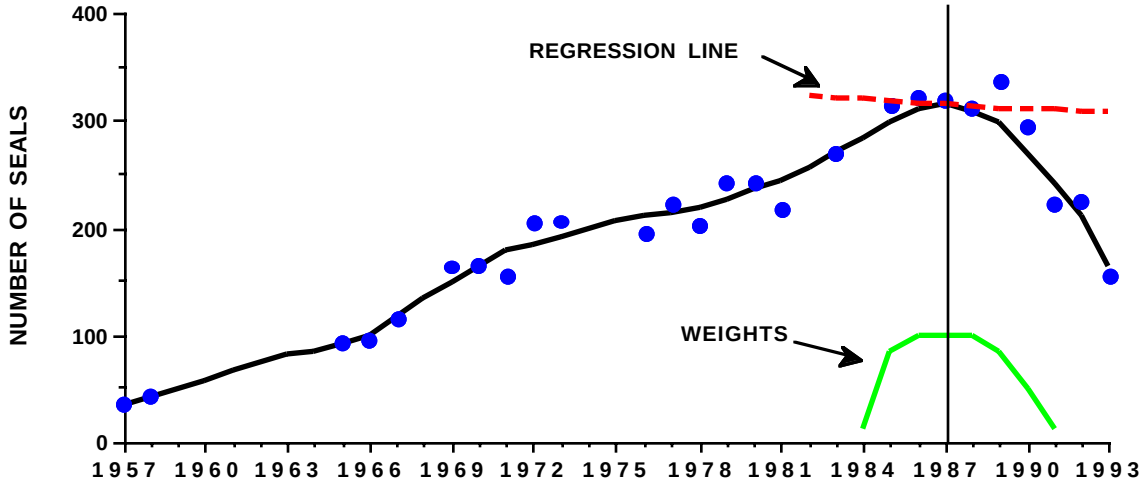


Fig. 1.18. Locally-weighted regression line ("lowess" smoothing) for the French Frigate Shoals monk seal beach counts. For each year on the graph, a weighted linear regression is computed from the n nearest points, with the contribution of each point weighted by a cubic function of distance of the data point from the base point. The regression line for 1987 is shown on the graph, along with the weights assigned to the 9 nearest points. The regression line determines only one point on the smoothed line.

1.15 Non-linear least-squares

The method of least-squares was discussed in Sec. 1.4, and eq.(1.6) was used to develop least-squares estimates for linear regression. The same approach can be used to fit non-linear functions, starting with the same equation for sum of squares:

$$S = \Sigma[y_i - f(x)]^2$$

where f(x) is now some non-linear function, such as the logistic function of eq. (1.30). One could find a minimum for the sum of squares, S, by a direct search routine. This is labor-intensive, and there are various computer programs that do the job very quickly and efficiently. Some of these call for partial derivatives of the function (used to "linearize" the function so that the approach to a minimum can be done in successive iterations). Others use numerical approximations to the partial derivatives, or direct search routines. SYSTAT contains two such routines under the "nonlin" function. It requires that a model be furnished, but this can be written in the notation used in EXCEL (really statements in BASIC language, which underlies EXCEL). Thus eq. (1.30) is entered as:

$$P = 1/(1 + EXP(-(A+B*X)))$$

The data need to be entered by using the Editor function (or can be read in from an EXCEL file, or copied to the Editor via a clipboard). Names used for variables (P,X) above are used as column headers in the data file, and the SYSTAT fitting routine recognizes the other labels (except built-in functions like EXP) as variables to fit (such as A and B above). Trial values can be furnished (i.e., rough estimates of A and B) and the number of iterations can be set (these have built-in "default" values). It may be necessary to use trial values if the program doesn't converge in, say 20 iterations (the default value)), but further iterations can be tried, first. Since the program is iterative, it may get stuck in various ways, and it is then desirable to quit, and start over with different guesses at starting parameters.

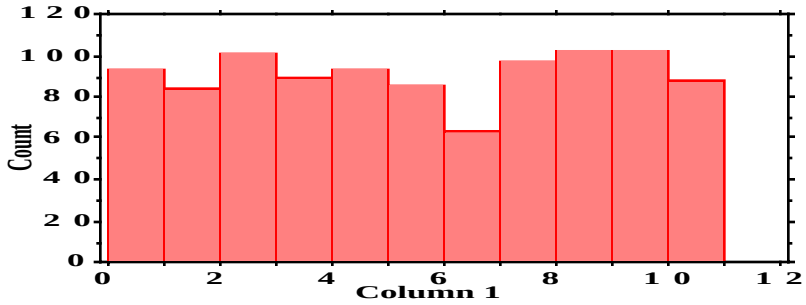
1.16 Exercises

1.16.1 Coin-tossing. Students should try a coin-tossing experiment like the one reported in example 1.1. Put 10 coins in a jar and make 100 tosses, recording the number of heads in blocks of 10. Make a frequency distribution and compare it with Fig. 1.1. Try another set of 100 and compare the two frequency distributions. Compute the sample means and variances, and compare them with the theoretical values.

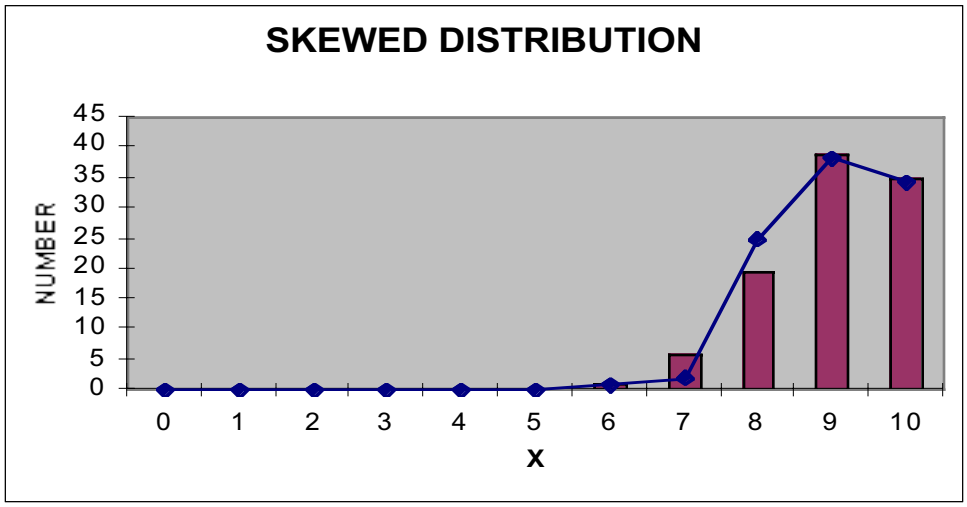
1.16.2 Simulating the binomial on a computer. Coin-tossing gets tiresome after awhile, and it is important to look at a different probability model. In order to get large samples without the tedium of mechanical approaches, we can resort to the computer. Students familiar with a programming language will likely prefer to write a simple program. However, useful results can be obtained in EXCEL and are readily in reach of those without programming experience. Those with only a passing experience with EXCEL may have to resort to the HELP function (or a colleague with experience) but it is important to carry out the following exercise because it should provide a capsule view of “monte carlo” simulations. Also, the next two chapters on bootstrapping depend on use of EXCEL. Insert the statement “=RANDBETWEEN(0,1)” in a cell in an EXCEL spreadsheet, and copy down to fill 10 cells in a column. This generates a series of 0’s and 1’s with probability 1/2 of getting either. Now copy the row to the right for 100 columns (it is convenient to use the automatic numbering system in a column above the 10 entries to keep track—a handy little number pops up beside to indicate how many numbers you have entered). Now sum the columns (use the summation function in the legend at the top of the sheet). This row of numbers (the sums) is now equivalent to the table of data in Example 1.1. Now use the histogram procedure (in the Tools menu) to construct

a histogram of frequencies of results. These should approximate the bars in Fig. 1.1. Note that every time you make a change in the worksheet it recalculates the table of random values (this function can be turned off). It is worthwhile to calculate several histograms just to get a notion of how variable the outcomes are. Next calculate the expected values from eq.(1.1). Find the factorial function ("FACT" in Math and Trig functions). Actually, all you need to know is that FACT(5) gives the value of 5! Use this function to calculate the factorial part of eq.(1.1) next to a column numbered 0 to 10. Then enter the rest of the equation in the next column (because p=0.5=1-p these entries will all be the same, but we'll use the approach for a case where p is not 1/2 below). The product of the two columns gives the proportions of eq.(1.1) which add to unity. Now multiply 100 times the proportions, and you have the expected values, which should approximate what you have in the histograms. The Chart Wizard in EXCEL will plot expected and observed values (you need to look under "Custom Types" to find one that plots a line and bars). One last chore is to recalculate the expected values using a value of p=0.9 which gives a distinctly asymmetric graph. It is always useful to put the numerical value of p above the calculations and use the "$" (e.g.,$A$30) notation to denote p in calculations for the equation. This lets one experiment with different values of p. Students should save a worksheet with the above calculations in order to have it for further reference when we consider other frequency distributions.

1.16.3 Random sampling There will be a great deal of emphasis on random sampling in this course. A relatively new topic in statistical methodology called bootstrapping will be used extensively. It depends on random sampling with replacement. Courses and books on sampling methodology usually depend on sampling without replacement. Consider using a number of sample plots to make counts of plants in order to estimate overall density of some species of plant. Such plots should be located at random in order to assure an unbiased estimate of density, and secure a reliable estimate of variance. Ordinarily, an investigator would find some way to assign a number to all possible plots in the area to be studied, and locate the sample plots by consulting a table of random numbers. If the same plot is drawn twice, it would not be counted twice, as this usually makes no sense. Hence we describe this as sampling without replacement. Textbooks on sampling show that it usually doesn't make much difference whether we do in fact sample with replacement, inasmuch as as the sample usually is a small fraction of the total population. Bootstrapping, however, depends on sampling with replacement as a way to reflect the underlying frequency distribution. Consequently, most of our samples will be with replacement. We will be taking repeated random samples with replacement of a data set. The individual entries in the data set will be in a computer file, and we will randomly select individual entries from this file. It is convenient to number the data items from 1 to n, and we then need to generate random numbers. To illustrate the approach, enter "RANDBETWEEN(1,10)" in a cell in EXCEL and copy down the column for 100 entries. Make a histogram of the data, as in Example 1.16.2. This is a sample from a uniform distribution, i.e., a frequency distribution where the probabilities are all equal. It is the distribution underlying random sampling. It is easy to extend the process to, say, 1,000 draws as in the frequency distribution plotted below. Note that it is still quite variable, even with 1,000 draws. Make a graph of your data like the following using the Chart Wizard and post it on a spreadsheet with the calculations.



1.16.4 Simulating a discrete skewed distribution. In Exercise 1.16.2 students were asked to calculate expected values for a binomial frequency distribution [eq.(1.1)] with p=0.9. A skewed frequency distribution is not hard to simulate, requiring two changes to the methods used in Exercise 1.16.2. Instead of RANDBETWEEN(0,1) we use "=RAND()" (don't put anything in the parenthes) which provides random numbers between 0 and 1. We also need an "IF" function which is the basis for a lot of computer work. It evaluates an expression and chooses between two output values, depending on whether the expression is true or false (there are a number of different expressions working along these lines, but we use the simplest here). Set up a spreadsheet with a column of 10 values of "=RAND()", and copy it to the right 100 times. We again need a numerical value of p above this table for reference, which may be say 0.9. If the first entry in the first column is in position, say, D9, then in the column just below this first column the first entry should be "=If(D9>=A3,1,0)" where the value of p is in A3. Copy down 10 and across for 100 columns and sum these entries. The IF function checks to see if the entry in D9 exceeds p and enters 1 if true and 0 if false. The sums then provide the basis for a histogram of a skewed discrete distribution. Make histograms with p=0.1, and p=0.5. Compare the histogram with p=0.5 with the one you made in Exercise 1.16.2. Make a new calculation of eq.(1.1) with p=0.9 and compare it with the histogram with p=0.1 (actually you should have made one in Exercise 16.1.2 and need only copy it over to this worksheet for comparison. Plot the data in Chart Wizard (expected and observed values). It should look like the following graph:



1.16.5. Do the algebra to calculate the expected value of eq.(1.1) as given in the right side of eq.(1.2).

1.16.6 Simulating a continuous skewed distribution. A continuous random variable is one that has values in the real domain. For our purposes, this means values like those generated by RAND() -- any number within the range considered (i.e., from 0 to 1). We will consider one way to generate random variables from an exponential distribution here. Consider the function:

$$F(x) = 1 - e^{-\beta x} \qquad (1.34)$$

This is an example of a cumulative distribution or cumulative distribution function. It takes values from 0 to 1, and has one parameter, β, which controls the rate at which the function approaches unity. The graph below shows the function for β = 0.1 and β = 0.5.

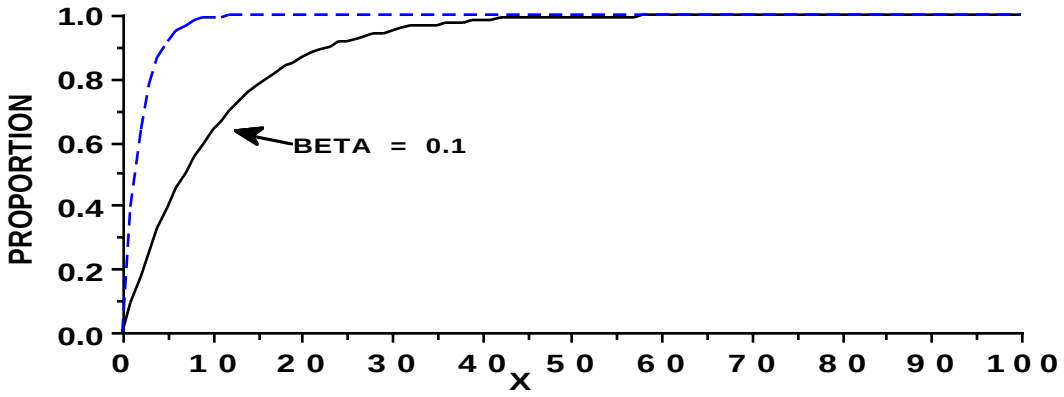


Fig. 1.19 Plot of cumulative distribution function for the exponential distribution for values of β = 0.1 (solid line) and β = 0.5 (broken line).

We use the cumulative distribution function here because it takes values from 0 to 1, and we can take a random sample from that range (using RAND()) and translate that to find the corresponding x value, by rearranging eq. (1.34) as

$$x = \frac{1}{\beta}\log_e(1 - F(x)) \qquad (1.35)$$

Thus the procedure is to draw a random sample of values from RAND() and look up the corresponding values of x. Eq. (1.34) is the integral of an exponential distribution over the range 0 to x, hence the name "cumulative". To compare the outcomes of a simulation with the equation for the frequency distribution, one runs a simulation as described in Exercise 1.16.4, and plots the results. Differentiating the cumulative yields the frequency distribution:

$$f(x) = \frac{dF(x)}{dx} = \frac{d[1 - e^{-\beta x}]}{dx} = \beta e^{-\beta x} \qquad (1.36)$$

Students whose calculus is a little rusty may want to look up the formula for finding a differential of an exponential; others may want to accept the statement without derivation. We need the right side of eq.(1.36) only to be able to compare simulation outcomes with the theoretical model, given in the figure below. Produce a column of 1000 random variables [F(x)] from RAND() and convert them with eq.(1.35), make a histogram of the results (using 30 “bins”) and then calculate the expected values by multiplying eq.(1.36) times 1000. Plot these as before and see how your result compares with the graph below.

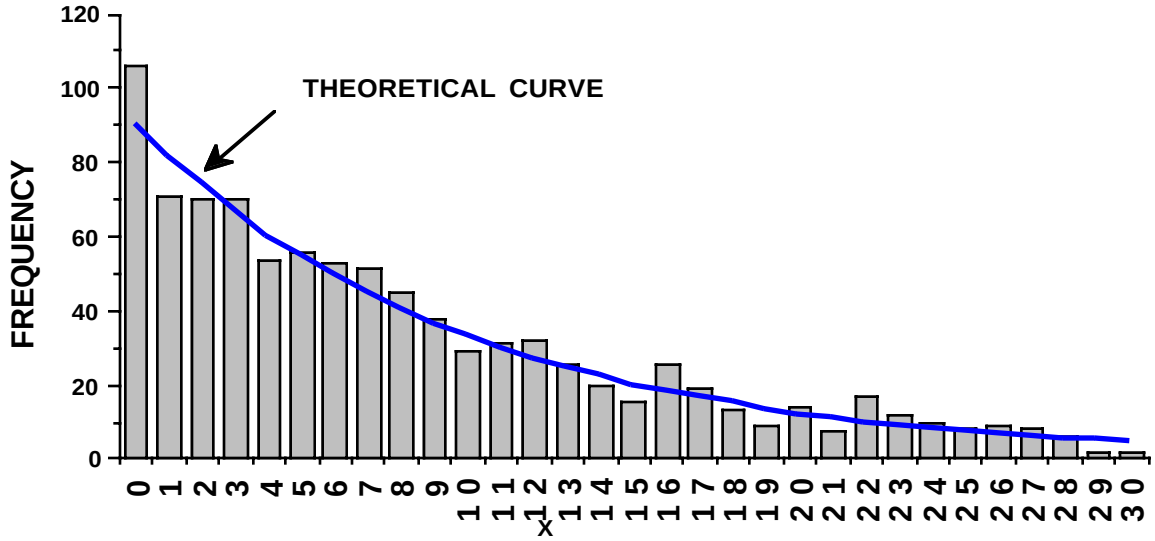


Fig. 1.20. Simulated exponential data compared to theoretical curve.

1.16.7 Simple linear regression. Data on counts of deer on a study area are given below. Fit the linear regression of Fig. 1.3 by using eqs. (1.9). This is readily done in EXCEL (in fact, EXCEL has a regression fitting routine which we will use for additional exercises, but students should do the calculations directly from the definitions in order to see how they "work" and then check by using the built-in fitting routine). Some graphics programs will also do the fitting automatically.

Year	Number of deer
x_i	y_I
1	10
2	21
3	52
4	71
5	97
6	146
7	212

1.16.8 Check the fact that a and b give minimum values of eq.(1.6), the sum of squares, for the deer data of Exercise 1.16.7. Copy the results of Exercise 1.16.7 into a new worksheet and compute eq. (1.6) for a and b, setting up the worksheet so that a and b are listed as separate entries on the worksheet as shown below. Then vary a and b by small amounts and write down the resulting sums of squares in the table. That is, make a table like the following and fill in the entries. It is easiest to first make your entries in pencil as transferring them individually to a summary table in EXCEL calls for a lot of tedious use of "Paste Special" in the menu, and/or provides opportunities to forget which cell you were working with. You should find a minimum in this table. If you want to try to get closer to the values of a and b found in Exercise 1.16.6, make a new table with fractional values in the row and column headings (e.g.,31.1, 31.2, etc.) and fill in the new table. This approach provides a device that is sometimes useful to solve a pair of more complex equations without needing to use a non-linear least-squares fitting routine. It is tedious unless you can guess reliably in advance just which part of the "Sums-of-

Squares" space the answer lies. But the purpose here is just to show how things work.

Sums-of-Squares table(eq.1.6)

		b			
	30	31	32	33	34
	-39	2046			
	-40				
	-41				2520
a	-42		1987		
	-43				
	-44				
	-45				2184

1.16.9 Use EXCEL and eq. (1.11) to calculate ANOVA for a regression equation for the data of Exercise 1.16.7 and compare your results with those given in Table 1.1. Now use the EXCEL regression program (found in the same group of analysis tools as are the ANOVA programs) to see how it works, and add the results to your direct computations.

1.16.10 Compute the correlation coefficient for the deer data from eq. (1.12). It can also be directly computed using a function CORREL found in the functions menu.

1.16.11 Compute s^2_b from eq. (1.15) for the deer data. Now compute it assuming that you have 3 observations (9,10,11) from year 1 and 4 observations (207,210,212, 219) at year 7 (and no observations for years 2,3,4,5 and 6). You will need to recalculate everything for the new data. What do you conclude about the effect of this arrangement of the data on s^2_b? Would you recommend this approach? Why?

1.16.12 Compute confidence limits for b from eq.(1.16) using the following set of data. Show details of your computation (i.e., the components of the calculation on a spreadsheet).

1	2.86
2	0.90
3	1.56
4	3.85
5	1.62
6	4.39
7	3.66
8	3.95
9	4.45
10	4.50

Note that the α in eq.(1.16) is not the same as α in the regression model. It is standard notation for the probability level. Use α= 0.05 here. You can obtain the needed t-value from the functions in EXCEL (f_x on the Toolbar) which is TINV(α,d.f.) where α is the desired probability for a 2-tailed t-test. You can run the regression analysis in EXCEL to confirm your results.

1.16.13 Multiple regression. Calculate a multiple regression equation on the following data, using eqs.(1.19) and check your results in EXCEL. The data were used in an early effort to construct an index of abundance for grizzly bears in Yellowstone National Park. Use the logarithm of the count as y and "Yr." As x1

and "Freq. Sight" as x2. It is important not to use the actual 4 digit year as x1 because it can cause a loss of accuracy when larger data sets are involved.

Year	Count	ln count	Yr.	Freq. sight.
1976	17	2.8332	1	1.64
1977	13	2.5649	2	1.50
1978	9	2.1972	3	1.28
1979	13	2.5649	4	1.08
1980	12	2.4849	5	1.40
1981	14	2.6391	6	1.58
1982	11	2.3979	7	1.62
1983	13	2.5649	8	1.20
1984	17	2.8332	9	2.29
1985	9	2.1972	10	2.00
1986	25	3.2189	11	3.12
1987	13	2.5649	12	1.64
1988	19	2.9444	13	2.12
1989	16	2.7726	14	1.86
1990	25	3.2189	15	1.95
1991	24	3.1781	16	2.65
1992	23	3.1355	17	1.65
1993	20	2.9957	18	1.67
1994	20	2.9957	19	1.47

1.16.14 Perform the test for curvilinearity described in the text (Sec. 1.7) and illustrated in Table 1.5 on the following sets of data. Make a spreadsheet containing the ANOVA tables (as in Table 1.5); note that the deer data are also included here so you have an example of the expected results at hand) and discuss the results as they apply to the notion that one should test for the assumptions before doing an analysis. Do the tests of the ANOVA tables provide convincing evidence of nonlinearity in the data?

Year	Horses	Year	Deer	Year	Elk
1	340	1	10	1	3172
2	423	2	21	2	4305
3	482	3	52	3	5543
4	611	4	71	4	7281
5	762	5	97	5	8215
6	879	6	146	6	9981
		7	212	7	10529
				8	12607

Year	Gray seals	Year	Muskox
1	751	1	49
2	854	2	57
3	869	3	65
4	898	4	61
5	1019	5	76

1.16.15 The following data are replicate monk seal beach counts from French Frigate Shoals. Conduct a test for significant deviations from regression using the "pure error" model of Section 1.6. There may be an advantage in using logarithms of the counts (to approximately "normalize" the data), as was done

in Section 1.6, but try the analysis without the log transform. Report your results in an analysis of variance on a spreadsheet, as in Table 1.3.

1985	298	1990	264	1994	193
	250		271		183
	301		262		219
	403		300		190
1986	401		299		196
	285		300		198
	278	1991	176		202
1987	351		191		232
	285		216		222
	316		217		249
	301		197	1995	141
	320		185		124
	350		281		168
	333		273		132
	252	1992	204		140
	362		202		144
1988	292		226		174
	303		227		156
	288		234		164
	286		271		
	315		231		
	327	1993	156		
	327		195		
	354		186		
1989	331		182		
	337		189		
	322		221		
	313		161		
	279		184		
	292		187		
	319		208		
	354		194		
	375		219		
	363				

1.16.16 The following data are from three years of a survey of harbor porpoises in which there were replicate transects flown and the transect lengths were recorded.

Year	Km.	Count		Km.	Count		Km.	Count
1986	552	48	1987	326.6	1	1988	199.1	1
	318	31		117.5	12		66.5	12
	445	9		752	30		374.7	5
	399	59		384.4	24		685.7	71
	195	1		58.5	0		333.7	0
	150	10		223.2	6		311.9	18

Test for significant differences in slopes of regressions of porpoise counts on transect lengths as done in Section 1.9. Report your results on a spreadsheet.

1.16.17 Perform an analysis of covariance (Section 1.10) on the data of Exercise 1.16.16, and report your results on a spreadsheet. Is this a legitimate analysis in view of the results of Exercise 1.16.16? Explain.